

# **Australasian Language Technology Association Workshop 2018**

**Proceedings of the Workshop**



**Editors:**  
**Sunghwan Mac Kim**  
**Xiuzhen (Jenny) Zhang**

**10–12 December 2018**  
**The University of Otago**  
**Dunedin, New Zealand**

**Australasian Language Technology Association Workshop 2018**  
**(ALTA 2018)**

<http://alta2018.alta.asn.au>

Online Proceedings:  
<http://alta2018.alta.asn.au/proceedings>

**Gold Sponsors:**



**Silver Sponsor:**



**Bronze Sponsors:**



# **ALTA 2018 Workshop Committees**

## **Workshop Co-Chairs**

- Xiuzhen (Jenny) Zhang (RMIT University)
- Sunghwan Mac Kim (CSIRO Data61)

## **Publication Co-chairs**

- Sunghwan Mac Kim (CSIRO Data61)
- Xiuzhen (Jenny) Zhang (RMIT University)

## **Programme Committee**

- Abeed Sarker, University of Pennsylvania
- Alistair Knott, University of Otago
- Andrea Schalley, Karlstads Universitet
- Ben Hachey, The University of Sydney and Digital Health CRC
- Benjamin Borschinger, Google
- Bo Han, Accenture
- Daniel Angus, University of Queensland
- Diego Mollá, Macquarie University
- Dominique Estival, Western Sydney University
- Gabriela Ferraro, CSIRO Data61
- Gholamreza Haffari, Monash University
- Hamed Hassanzadeh, Australian e-Health Research Centre CSIRO
- Hanna Suominen, Australian National University and Data61/CSIRO
- Jennifer Biggs, Defence Science Technology
- Jey-Han Lau, IBM Research
- Jojo Wong, Monash University
- Karin Verspoor, The University of Melbourne
- Kristin Stock, Massey University
- Laurianne Sitbon, Queensland University of Technology
- Lawrence Cavedon, RMIT University
- Lizhen Qu, CSIRO Data61
- Long Duong, Voicebox Technology Australia
- Maria Kim, Defence Science Technology
- Massimo Piccardi, University of Technology Sydney
- Ming Zhou, Microsoft Research Asia
- Nitin Indurkhy, University of New South Wales
- Rolf Schwitter, Macquarie University
- Sarvnaz Karimi, CSIRO Data61
- Scott Nowson, Accenture
- Shervin Malmasi, Macquarie University and Harvard Medical School
- Shunichi Ishihara, Australian National University

- Stephen Wan, CSIRO Data61
- Teresa Lynn, Dublin City University
- Timothy Baldwin, The University of Melbourne
- Wei Gao, Victoria University of Wellington
- Wei Liu, University of Western Australia
- Will Radford, Canva
- Wray Buntine, Monash University

## Preface

This volume contains the papers accepted for presentation at the Australasian Language Technology Association Workshop (ALTA) 2018, held at The University of Otago in Dunedin, New Zealand on 10-12 December 2018.

The goals of the workshop are to:

- bring together the Language Technology (LT) community in the Australasian region and encourage interactions and collaboration;
- foster interaction between academic and industrial researchers, to encourage dissemination of research results;
- provide a forum for students and young researchers to present their research;
- facilitate the discussion of new and ongoing research and projects;
- increase visibility of LT research in Australasia and overseas and encourage interactions with the wider international LT community.

This year's ALTA Workshop presents 10 peer-reviewed papers, including 6 long papers and 4 short papers. We received a total of 17 submissions for long and short papers. Each paper was reviewed by three members of the program committee, using a double-blind protocol. Great care was taken to avoid all conflicts of interest.

ALTA 2018 includes a presentations track, following the workshops since 2015 when it was first introduced. This aims to encourage broader participation and facilitate local socialisation of international results, including work in progress and work submitted or published elsewhere. Presentations were lightly reviewed by the ALTA chairs to gauge overall quality of work and whether it would be of interest to the ALTA community. Offering both archival and presentation tracks allows us to grow the standard of work at ALTA, to better showcase the excellent research being done locally.

ALTA 2018 continues the tradition of including a shared task, this year on classifying patent applications. Participation is summarised in an overview paper by organisers Diego Mollá and Dilesha Seneviratne. Participants were invited to submit a system description paper, which are included in this volume without review.

We would like to thank, in no particular order: all of the authors who submitted papers; the programme committee for the time and effort they put into maintaining the high standards of our reviewing process; the shared task organisers Diego Mollá and Dilesha Seneviratne; our keynote speakers Alistair Knott and Kristin Stock for agreeing to share their perspectives on the state of the field; and the tutorial presenter Phil Cohen for his efforts towards the tutorial of collaborative dialogue. We would like to acknowledge the constant support and advice of the ALTA Executive Committee such as budgets, sponsorship and more.

Finally, we gratefully recognise our sponsors: CSIRO/Data61, Soul Machines, Google, IBM, Seek and ARC Centre of Excellence for the Dynamics of Language. Importantly, their generous support enabled us to offer travel subsidies to all students presenting at ALTA, and helped to subsidise conference catering costs and student paper awards.

Xiuzhen (Jenny) Zhang  
Sunghwan Mac Kim  
ALTA Programme Chairs

# **ALTA 2018 Programme**

## **Monday, 10 December 2018**

---

Tutorial Session (Room 1.19)

14:00–17:00	Tutorial: Phil Cohen <i>Towards Collaborative Dialogue</i>
17:00	End of Tutorial

---

## Tuesday, 11 December 2018

---

Opening & Keynote (Room 1.17)

9:00–9:15

Opening

9:15–10:15

Keynote 1 (from ADCS): Jon Degenhardt (Room 1.17)

*An Industry Perspective on Search and Search Applications*

---

10:15–10:45

Morning tea

Session A: Text Mining & Applications (Room 1.19)

10:45–11:05

Paper: Rolando Coto Solano, Sally Akevai Nicholas and Samantha Wray

*Development of Natural Language Processing Tools for Cook Islands Māori*

11:05–11:25

Paper: Bayzid Ashik Hossain and Rolf Schwitter

*Specifying Conceptual Models Using Restricted Natural Language*

11:25–11:45

Presentation: Jenny McDonald and Adon Moskal

*Quantext: a text analysis tool for teachers*

11:45–11:55

Paper: Xuanli He, Quan Tran, William Havard, Laurent Besacier, Ingrid Zukerman and Ghoreza Haffari

*Exploring Textual and Speech information in Dialogue Act Classification with Speaker Domain Adaptation*

---

11:55–13:00

Lunch

13:00–14:00

Keynote 2: Alistair Knott (Room 1.17)

*Learning to talk like a baby*

---

14:00–14:15

Break

14:15–15:15

Poster Session 1 (ALTA & ADCS)

---

15:15–15:45

Afternoon tea

15:45–16:50

Poster Session 2 (ALTA & ADCS)

---

16:50

End of Day 1

---

## **Wednesday, 12 December 2018**

9:00–10:15	Keynote 3 (from ADCS): David Bainbridge (Room 1.17) <i>Can You Really Do That? Exploring new ways to interact with Web content and the desktop</i>
10:15–10:45	Morning tea
	Session B: Machine Translation & Speech (Room 1.19)
10:45–11:05	Paper: Cong Duy Vu Hoang, Gholamreza Haffari and Trevor Cohn <i>Improved Neural Machine Translation using Side Information</i>
11:05–11:25	Presentation: Qiongkai Xu, Lizhen Qu and Jiawei Wang <i>Decoupling Stylistic Language Generation</i>
11:25–11:45	Paper: Satoru Tsuge and Shunichi Ishihara <i>Text-dependent Forensic Voice Comparison: Likelihood Ratio Estimation with the Hidden Markov Model (HMM) and Gaussian Mixture Model – Universal Background Model (GMM-UBM) Approaches</i>
11:45–11:55	Paper: Nitika Mathur, Timothy Baldwin and Trevor Cohn <i>Towards Efficient Machine Translation Evaluation by Modelling Annotators</i>
11:55–12:55	Lunch
12:55–13:55	Keynote 4: Kristin Stock (Room 1.17) <i>"Where am I, and what am I doing here?" Extracting geographic information from natural language text</i>
13:55–14:05	Break
	Session C: Shared session with ADCS (Room 1.17)
14:05–14:30	Paper: Alfan Farizki Wicaksono and Alistair Moffat (ADCS long paper) <i>Exploring Interaction Patterns in Job Search</i>
14:30–14:50	Paper: Xavier Holt and Andrew Chisholm <i>Extracting structured data from invoices</i>
14:50–15:05	Paper: Bevan Koopman, Anthony Nguyen, Danica Cossio, Mary-Jane Courage and Gary Francois (ADCS short paper) <i>Extracting Cancer Mortality Statistics from Free-text Death Certificates: A View from the Trenches</i>
15:05–15:15	Paper: Hanieh Poostchi and Massimo Piccardi <i>Cluster Labeling by Word Embeddings and WordNet's Hypernymy</i>
15:15–15:35	Afternoon tea
	Session D: Word Semantics (Room 1.19)
15:35–15:55	Paper: Lance De Vine, Shlomo Geva and Peter Bruza <i>Unsupervised Mining of Analogical Frames by Constraint Satisfaction</i>
15:55–16:05	Paper: Navnita Nandakumar, Bahar Salehi and Timothy Baldwin <i>A Comparative Study of Embedding Models in Predicting the Compositionality of Multiword Expressions</i>
	Shared Task Session (Room 1.19)
16:05–16:15	Paper: Diego Mollá-Aliod and Dilesha Seneviratne <i>Overview of the 2018 ALTA Shared Task: Classifying Patent Applications</i>
16:15–16:25	Paper: Fernando Benites, Shervin Malmasi, Marcos Zampieri <i>Classifying Patent Applications with Ensemble Methods</i>
16:25–16:35	Paper: Jason Hepburn <i>Universal Language Model Fine-tuning for Patent Classification</i>
16:35–16:45	Break
16:45–16:55	Best Paper Awards
16:55–17:20	Business Meeting & ALTA Closing
17:20	End of Day 2

## Contents

<b>Invited talks</b>	<b>1</b>
<b>Tutorials</b>	<b>3</b>
<b>Long papers</b>	<b>5</b>
<i>Improved Neural Machine Translation using Side Information</i>	
<b>Cong Duy Vu Hoang, Gholamreza Haffari and Trevor Cohn</b>	<b>6</b>
<i>Text-dependent Forensic Voice Comparison: Likelihood Ratio Estimation with the Hidden Markov Model (HMM) and Gaussian Mixture Model</i>	
<b>Satoru Tsuge and Shunichi Ishihara</b>	<b>17</b>
<i>Development of Natural Language Processing Tools for Cook Islands Māori</i>	
<b>Rolando Coto Solano, Sally Akevai Nicholas and Samantha Wray</b>	<b>26</b>
<i>Unsupervised Mining of Analogical Frames by Constraint Satisfaction</i>	
<b>Lance De Vine, Shlomo Geva and Peter Bruza</b>	<b>34</b>
<i>Specifying Conceptual Models Using Restricted Natural Language</i>	
<b>Bayzid Ashik Hossain and Rolf Schwitter</b>	<b>44</b>
<i>Extracting structured data from invoices</i>	
<b>Xavier Holt and Andrew Chisholm</b>	<b>53</b>
<b>Short papers</b>	<b>60</b>
<i>Exploring Textual and Speech information in Dialogue Act Classification with Speaker Domain Adaptation</i>	
<b>Xuanli He, Quan Tran, William Havard, Laurent Besacier, Ingrid Zukerman and Gholamreza Haffari</b>	<b>61</b>
<i>Cluster Labeling by Word Embeddings and WordNet's Hypernymy</i>	
<b>Hanieh Poostchi and Massimo Piccardi</b>	<b>66</b>
<i>A Comparative Study of Embedding Models in Predicting the Compositionality of Multiword Expressions</i>	
<b>Navnita Nandakumar, Bahar Salehi and Timothy Baldwin</b>	<b>71</b>
<i>Towards Efficient Machine Translation Evaluation by Modelling Annotators</i>	
<b>Nitika Mathur, Timothy Baldwin and Trevor Cohn</b>	<b>77</b>

<b>ALTA Shared Task papers</b>	<b>83</b>
<i>Overview of the 2018 ALTA Shared Task: Classifying Patent Applications</i>	
<b>Diego Mollá and Dilesha Seneviratne</b>	<b>84</b>
<i>Classifying Patent Applications with Ensemble Methods</i>	
<b>Fernando Benites, Shervin Malmasi, Marcos Zampieri</b>	<b>89</b>
<i>Universal Language Model Fine-tuning for Patent Classification</i>	
<b>Jason Hepburn</b>	<b>93</b>

## **Invited keynotes**

## **Alistair Knott (University of Otago & Soul Machines)**

### *Learning to talk like a baby*

In recent years, computational linguists have embraced neural network models, and the vector-based representations of words and meanings they use. But while computational linguists have readily adopted the machinery of neural network models, they have been slower to embrace the original aim of neural network research, which was to understand how brains work. A large community of neural network researchers continues to pursue this “cognitive modelling” aim, with very interesting results. But the work of these more cognitively minded modellers has not yet percolated deeply into computational linguistics. In my talk, I will argue the cognitive modelling tradition of neural networks has much to offer computational linguistics. I will outline a research programme that situates language modelling in a broader cognitive context. The programme is distinctive in two ways. Firstly, the initial object of study is a baby, rather than an adult. Computational linguistics models typically aim to reproduce adult linguistic competence in a single training process, that presents an “empty” network with a corpus of mature language. I will argue that this training process doesn’t correspond to anything in human experience, and that we should instead aim to model a more gradual developmental process, that first achieves babylike language, then childlike language, and so on. Secondly, the new programme studies the baby’s language system as it interfaces with her other cognitive systems, rather than by itself. It pays particular attention to the sensory and motor systems through which a baby engages with the physical world, which are the primary means by which it activates semantic representations. I will argue that the structure of these sensorimotor systems, as expressed in neural network models, offer interesting insights about certain aspects of linguistic structure. I will conclude by demoing a model of the interface between language and the sensorimotor system, as it operates in a baby at an early stage of language learning.

## **Kristin Stock (Massey University)**

### *“Where am I, and what am I doing here?” Extracting geographic information from natural language text*

The extraction of place names (toponyms) from natural language text has received a lot of attention in recent years, but location is frequently described in more complex ways, often using other objects as reference points. Examples include: ‘The accident occurred opposite the Orewa Post Office, near the pedestrian crossing’ or ‘the sample was collected on the west bank of the Waikato River, about 3km upstream from Huntly’. These expressions can be vague, imprecise, underspecified, rely on access to information about other objects in the environment, and the semantics of spatial relations like ‘opposite’ and ‘on’ are still far from clear. Furthermore, many of these kinds of expressions are context sensitive, and aspects such as scale, geometry and type of geographic feature may influence the way the expression is understood. Both machine learning and rule-based approaches have been developed to try to firstly parse expressions of this kind, and secondly to determine the geographic location that the expression refers to. Several relevant projects will be discussed, including the development of a semantic rather than syntactic approach to parsing geographic location descriptions; the creation of a manually annotated training set of geographic language; the challenges highlighted from human descriptions of location in the emergency services context; the interpretation and geocoding of descriptions of flora and fauna specimen collections; the development of models of spatial relations using social media data and the use of instance-based learning to interpret complex location descriptions.

## **Tutorials**

## Towards Collaborative Dialogue

*Phil Cohen (Monash University)*

This tutorial will discuss a program of research for building collaborative dialogue systems, which are a core part of virtual assistants. I will briefly discuss the strengths and limitations of current approaches to dialogue, including neural network-based and slot-filling approaches, but then concentrate on approaches that treat conversation as planned collaborative behaviour. Collaborative interaction involves recognizing someone's goals, intentions, and plans, and then performing actions to facilitate them. People have learned this basic capability at a very young age and are expected to be helpful as part of ordinary social interaction. In general, people's plans involve both speech acts (such as requests, questions, confirmations, etc.) and physical acts. When collaborative behavior is applied to speech acts, people infer the reasons behind their interlocutor's utterances and attempt to ensure their success. Such reasoning is apparent when an information agent answers the question "Do you know where the Sydney flight leaves?" with "Yes, Gate 8, and it's running 20 minutes late." It is also apparent when one asks "where is the nearest petrol station?" and the interlocutor answers "2 kilometers to your right" even though it is not the closest, but rather the closest one that is open. In this latter case, the respondent has inferred that you want to buy petrol, not just to know the location of the station. In both cases, the literal and truthful answer is not cooperative. In order to build systems that collaborate with humans or other artificial agents, a system needs components for planning, plan recognition, and for reasoning about agents' mental states (beliefs, desires, goals, intentions, obligations, etc.).

In this tutorial, I will discuss current theory and practice of such collaborative belief-desire-intention architectures, and demonstrate how they can form the basis for an advanced collaborative dialogue manager. In such an approach, systems reason about what they plan to say, and why the user said what s/he did. Because there is a plan standing behind the system's utterances, it is able to explain its reasoning. Finally, we will discuss potential methods for incorporating such a plan-based approach with machine-learned approaches.

## **Long papers**

# Improved Neural Machine Translation using Side Information

Cong Duy Vu Hoang<sup>†</sup> and Gholamreza Haffari<sup>‡</sup> and Trevor Cohn<sup>†</sup>

<sup>†</sup> University of Melbourne, Melbourne, VIC, Australia

<sup>‡</sup> Monash University, Clayton, VIC, Australia

vhoang2@student.unimelb.edu.au, gholamreza.haffari@monash.edu,  
t.cohn@unimelb.edu.au

## Abstract

In this work, we investigate whether side information is helpful in the context of neural machine translation (NMT). We study various kinds of side information, including topical information and personal traits, and then propose different ways of incorporating these information sources into existing NMT models. Our experimental results show the benefits of side information in improving the NMT models.

## 1 Introduction

Neural machine translation is the task of generating a target language sequence given a source language sequence, framed as a neural network (Sutskever et al., 2014; Bahdanau et al., 2015, *inter alia*). Most research efforts focus on inducing more prior knowledge (Cohn et al., 2016; Zhang et al., 2017; Mi et al., 2016, *inter alia*), incorporating linguistics factors (Hoang et al., 2016b; Sennrich and Haddow, 2016; García-Martínez et al., 2017) or changing the network architecture (Gehring et al., 2017b,a; Vaswani et al., 2017; El-bayad et al., 2018) in order to better exploit the source representation. Consider a different direction, situations in which there exists other modality other than the text of the source sentence. For instance, the WMT 2017 campaign<sup>1</sup> proposed to use additional information obtained from *images* to enrich the neural MT models, as in (Calixto et al., 2017; Matusov et al., 2017; Calixto and Liu, 2017). This task, also known as multi-modal translation, seeks to leverage images which can contain cues representing the perception of the image in source text, and potentially can contribute to resolve ambiguity (e.g., lexical, gender),

vagueness, out-of-vocabulary terms, and topic rel-evancy.

Inspired from the idea of multi-modal translation, in our work, we propose the use of another modality, namely metadata or side information. Previously, Hoang et al. (2016a) have shown the usefulness of side information for neural language models. This work will investigate the potential usefulness of side information for NMT models. In our work, we target towards *unstructured and heterogeneous* side information which potentially can be found in practical applications. Specifically, we investigate different kinds of side information, including topic keywords, personality information and topic classification. Then we study different methods with minimal efforts for incorporating such side information into existing NMT models.

## 2 Machine Translation Data with Side Information

First, let's explore some realistic scenarios in which the side information is potentially useful for NMT.

**TED Talks** The TED Talks website<sup>2</sup> hosts technical videos from influential speakers around the world on various topics or domains, such as: education, business, science, technology, creativity, etc. Thanks to users' contributions, most of such videos are subtitled in multiple languages. Based on this website, Cettolo et al. (2012) created a parallel corpus for the MT research community. Inspired by this, Chen et al. (2016) further customised this dataset and included an additional sentence-level topic information.<sup>3</sup> We consider such topic information as side information. Fig-

<sup>1</sup><http://www.statmt.org/wmt17/multimodal-task.html>

<sup>2</sup><https://www.ted.com/talks>

<sup>3</sup><https://github.com/wenhuchen/iwslt-2015-de-en-topics>

ure 1 illustrates some examples of this dataset. As can be seen, the keywords (second column, treated as side information) contain additional contextual information that can provide complementary cues so as to better guide the translation process. Let’s take an example in Figure 1 (TED Video Id 172), the keyword “art” provides cues for words and phrases in target sequence such as: “place, design”; whereas the keyword “tech” refers to “Media Lab, computer science”.

**Personalised Europarl** For the second dataset, we evaluate our proposed idea in the context of personality-aware MT. Mirkin et al. (2015) explored whether translation preserves personality information (e.g., demographic and psychometric traits) in statistical MT (SMT); and further Rabinovich et al. (2017) found that personality information like author’s gender is an obvious signal in source text, but it is less clear in human and machine translated texts. As a result, they created a new dataset for personalised MT<sup>4</sup> partially based on the original Europarl. The personality such as author’s gender will be regarded as side information in our setup. An excerpt of this dataset is shown in Figure 2. As can be seen from the figure, there exist many kinds of side information pertaining to authors’ traits, including identification (ID, name), native language, gender, date of birth/age, and plenary session date. Here, we will focus on the “gender” trait and evaluate whether it can have any benefits in the context of NMT complementing the work of Rabinovich et al. (2017) attempted a similar idea as part of a SMT, rather than NMT, system.

**Patent MT Collection** Another interesting data is patent translation which includes rich side information. PatTR<sup>5</sup> is a sentence-parallel corpus which is a subset of the MAREC Patent Corpus (Wäschle and Riezler, 2012a). In general, PatTR contains millions of parallel sentences collected from all patent text sections (e.g., title, abstract, claims, description) in multiple languages (English, French, German) (Wäschle and Riezler, 2012b; Simianer and Riezler, 2013). An appealing feature of this corpus is that it provides a labelling at a sentence level, in the form of IPC (International Patent Classification) codes. The IPC

<sup>4</sup><http://cl.haifa.ac.il/projects/pmt/index.shtml>

<sup>5</sup><http://www.cl.uni-heidelberg.de/statnlpgroup/pattr/>

codes explicitly provide a hierarchical classification of patents according to various different areas of technology to which they pertain. This kind of side information can provide a useful signal for MT task – which has not yet been fully exploited. Figure 3 gives us an illustrating excerpt of this corpus. We can see that each of sentence pair in this corpus is associated with any number of IPC label(s) as well as other metadata, e.g., patent ID, patent family ID, publication date. In this work, we consider only the IPC labels. The full meaning of all IPC labels can be found on the official IPC website,<sup>6</sup> however we provide in Figure 3 the glossess for each referenced label. Note that those IPC labels form a WordNet style hierarchy (Fellbaum, 1998), and accordingly may be useful in many other deep models of NLP.

### 3 NMT with Side Information

We investigate different ways of incorporating side information into the NMT model(s).

#### 3.1 Encoding of Side Information

In this work, we propose the use of unstructured *heterogeneous* side information, which is often available in practical datasets. Due the heterogeneity of side information, our techniques are based on a bag-of-words (BOW) representation of the side information, an approach which was shown to be effective in our prior work (Hoang et al., 2016a). Each element of the side information (a label, or word) is embedded using a matrix  $\mathbf{W}_e^s \in \mathcal{R}^{H_s \times |V_s|}$ , where  $|V_s|$  is the vocabulary of side information and  $H_s$  the dimensionality of the hidden space. These embedding vectors are used for the input to several different neural architectures, which we now outline.

#### 3.2 NMT Model Formulation

Recall the general formulation of NMT (Sutskever et al., 2014; Bahdanau et al., 2015, *inter alia*) as a conditional language model in which the generation of target sequence is conditioned on the source sequence (Sutskever et al., 2014; Bahdanau et al., 2015, *inter alia*), formulated as:

$$\begin{aligned} \mathbf{y}_{t+1} &\sim p_{\Theta}(\mathbf{y}_{t+1}|\mathbf{y}_{<t}, \mathbf{x}) \\ &= \text{softmax}(\mathbf{f}_{\Theta}(\mathbf{y}_{<t}, \mathbf{x})); \end{aligned} \quad (1)$$

<sup>6</sup><http://www.wipo.int/classifications/ ipc/en/>

TED Video Id	Keywords	German Sentence	English Sentence
172	arts,tech	Aber das Media Lab ist ein interessanter Ort, und es ist wichtig für mich, denn ich studierte ursprünglich Computerwissenschaften und erst später in meinem Leben habe ich Design entdeckt.	But the Media Lab is an interesting place, and it's important to me because as a student, I was a computer science undergrad, and I discovered design later on in my life.
645	politics,issues,business	In anderen Worten, ich glaube, dass der französische Präsident Sarkozy recht hat, wenn er über eine Mittelmeer Union spricht.	So in other words, I believe that President Sarkozy of France is right when he talks about a Mediterranean union.
1193	recreation,arts,issues	Eine andere Welt tat sich ungefähr zu dieser Zeit auf: Auftreten und Tanzen.	Another world was opening up around this time: performance and dancing.
692	politics,arts,issues,env	Dieses Gebäude beinhaltet die Weltgrößte Kollektion von Sammlungen und Artefakten die der Rolle des USA im Kampf auf der Chinesischen Seite gedenken. In diesem langen Krieg -- die "liegenden Tiger".	This building contains the world's largest collection of documents and artifacts commemorating the U.S. role in fighting on the Chinese side in that long war -- the Flying Tigers.
1087	politics,education	Es erlaubt uns, Kunst, Biotechnologie, Software und all solch wunderbaren Dinge zu schaffen.	It allows us to do the art, the biotechnology, the software and all those magic things.
208	recreation,education,arts,issues	Ich liebe Bartóks Musik, so wie Herr Teszler, und er hatte wirklich jede Aufnahme Bartóks die es gab.	I love Bartók's music, as did Mr. Teszler, and he has virtually every recording of Bartók's music ever issued.

Fig. 1 An example with side information (e.g., keywords) for MT with TED Talks dataset.

<b>English Sentence:</b> Accordingly , I consider it essential that both the identification of cattle and the labelling of beef be introduced as quickly as possible on a compulsory basis .
<b>German Sentence:</b> Entsprechend halte ich es auch für notwendig , daß die Kennzeichnung möglichst schnell und verpflichtend eingeführt wird , und zwar für Rinder und für Rindfleisch .
<b>Meta Info:</b> EUROID="2209" NAME="Schierhuber" LANGUAGE="DE" GENDER="FEMALE" DATE_OF_BIRTH="31 May 1946" SESSION_DATE="97-02-19" AGE="50"
<b>English Sentence:</b> Can the Commission say that it will seek to have sugar declared a sensitive product ?
<b>German Sentence:</b> Kann die Kommission sagen , dass sie danach streben wird , Zucker zu einem sensiblen Produkt erklären zu lassen ?
<b>Meta Info:</b> EUROID="22861" NAME="Ó Neachtain (UEN)." LANGUAGE="EN" GENDER="MALE" DATE_OF_BIRTH="22 May 1947" SESSION_DATE="03-09-02" AGE="56"
<b>English Sentence:</b> For example , Brazil has huge concerns about the proposals because the poor and landless there will suffer if sugar production expands massively , as is predicted .
<b>German Sentence:</b> So geht beispielsweise Brasilien bezüglich der Vorschläge enorme Bedenken , denn wenn die Zuckerproduktion , wie vorhergesagt , massiv expandiert , wird das die Not der Armen und Landlosen dort noch verstärken .
<b>Meta Info:</b> EUROID="28115" NAME="McGuinness (PPE-DE )." LANGUAGE="EN" GENDER="FEMALE" DATE_OF_BIRTH="13 June 1959" SESSION_DATE="05-02-22" AGE="45"
<b>English Sentence:</b> The European citizens ' initiative should be seen as an opportunity to involve people more closely in the EU 's decision-making process .
<b>German Sentence:</b> Die Europäische Bürgerinitiative ist als Chance zu werten , um die Menschen stärker in den Entscheidungsprozess der EU miteinzubeziehen .
<b>Meta Info:</b> EUROID="96766" NAME="Ernst Strasser" LANGUAGE="DE" GENDER="MALE" DATE_OF_BIRTH="29 April 1956" SESSION_DATE="10-12-15-010" AGE="54"

Fig. 2 An example with side information (e.g., author's gender highlighted in red) for MT with personalised Europarl dataset.

<b>English Sentence:</b> In the case of the actual value coinciding with and/or deviating from the desired value input, the device emits audible signals.
<b>German Sentence:</b> Bei Übereinstimmung und/oder Abweichung des Istwertes von der Sollwerteingabe gibt die Vorrichtung akustische Signale ab.
<b>Meta Info:</b> EP-0017737-A1 6068117 19801029 G01D,G01P
<b>English Sentence:</b> The invention relates to a feed device for teeth (21) which forms part of a device for connecting a saw-blade base body to teeth (22, 23) that are subdivided into a metallic and a non-metallic material area (25, 26).
<b>German Sentence:</b> Eine Zahnzuführleinrichtung (21) ist Teil einer Vorrichtung zum Verbinden eines Sägeblattgrundkörpers mit Zähnen (22, 23), die in einen metallischen und einen nicht-metallischen Materialbereich (25, 26) unterteilt sind.
<b>Meta Info:</b> WO-2001002130-A1 7913309 20010111 B23K,B23D
G → PHYSICS G01 → MEASURING; TESTING G01D → MEASURING NOT SPECIALLY ADAPTED FOR A SPECIFIC VARIABLE; ARRANGEMENTS FOR MEASURING TWO OR MORE VARIABLES NOT COVERED BY A SINGLE OTHER SUBCLASS; TARIFF METERING APPARATUS; TRANSFERRING OR TRANSDUCING ARRANGEMENTS NOT SPECIALLY ADAPTED FOR A SPECIFIC VARIABLE; MEASURING OR TESTING NOT OTHERWISE PROVIDED FOR
G01P → MEASURING LINEAR OR ANGULAR SPEED, ACCELERATION, DECELERATION OR SHOCK; INDICATING PRESENCE OR ABSENCE OF MOVEMENT; INDICATING DIRECTION OF MOVEMENT
B → PERFORMING OPERATIONS; TRANSPORTING B23 → MACHINE TOOLS; METAL-WORKING NOT OTHERWISE PROVIDED FOR B23K → SOLDERING OR UNSOLDERING; WELDING; CLADDING OR PLATING BY SOLDERING OR WELDING; CUTTING BY APPLYING HEAT LOCALLY, e.g. FLAME CUTTING; WORKING BY LASER BEAM B23D → PLANING; SLOTTING; SHEARING; BROACHING; SAWING; SCRAPING; LIKE OPERATIONS FOR WORKING METAL BY REMOVING MATERIAL, NOT OTHERWISE PROVIDED FOR

Fig. 3 An example with side information (e.g., IPC highlighted in red) for MT with PatTR dataset.

where the probability  $p_{\Theta}(\cdot)$  of generating the next target word  $y_{t+1}$  is conditioned on the previously generated target words  $\mathbf{y}_{<t}$  and the source sequence  $\mathbf{x}$ ;  $f$  is a neural network which can be framed as an encoder-decoder model (Sutskever et al., 2014) and can use an attention mechanism (Bahdanau et al., 2015; Luong et al., 2015). In this model, the encoder encodes the information of the source sequence; whereas, the decoder decodes the target sequence sequentially from left-to-right. The attention mechanism controls which parts of the source sequence where the decoder should attend to in generating each symbol of target sequence. Later, advanced models have been proposed with modifications of the encoder and decoder architectures, e.g., using the 1D (Gehring et al., 2017b,a) and 2D (Elbayad et al., 2018) convolutions; or a transformer network (Vaswani et al., 2017). These advanced models have led to significantly better results in terms of both performance and efficiency via different benchmarks (Gehring et al., 2017b,a; Vaswani et al., 2017; Elbayad et al., 2018).

Regardless of the NMT architecture, we aim to explore in which case side information can be useful, as well as the effective and efficient way of incorporating them with minimal modification of the NMT architecture. Mathematically, we formulate the NMT problem given the availability of side information  $e$  as follows:

$$\begin{aligned} \mathbf{y}_{t+1} &\sim p_{\Theta}(y_{t+1} | \mathbf{y}_{<t}, \mathbf{x}, e) \\ &= \text{softmax}(f_{\Theta}(\mathbf{y}_{<t}, \mathbf{x}, e)); \end{aligned} \quad (2)$$

where  $e$  is the representation of additional side information we would like to incorporate into NMT model.

### 3.3 Conditioning on Side Information

Keeping in mind that we would like a generic incorporation method so that only minimal modification of NMT model is required, we propose and evaluate different approaches.

**Side Information as Source Prefix/Suffix** The most simple way to include side information is to add the side information as a string prefix or suffix to the source sequence, and letting the NMT model learn from this modified data. This method requires no modification of the NMT model. This method was firstly proposed by Sennrich et al. (2016a) who added the side constraints (e.g., hon-

orifics) as suffix of the source sequence for controlling the politeness in translated outputs.

**Side Information as Target Prefix** Alternatively, we can add the bag of words as a target prefix, inspired from Johnson et al. (2017) who introduces an artificial token as a prefix for specifying the required target language in a multilingual NMT system. Note that this method leads to additional benefits in the following situations: a) when the side information exists, the model takes them as inputs and then does its translation task as normal; b) when the side information is missing, so the model first generates the side information itself and subsequently uses it to proceed with translation.

**Output Layer** Similar to Hoang et al. (2016a) – who considers side information in the model focusing on the output side which worked well in LM, this method involves in two phases. First, it transforms the representation of the side information into a *summed* vector representation,  $e = \sum_{m \in [1, M]} e_{w_m^s}$ . We also tried the *average* operator in our preliminary experiments but observed no difference in end performance.

Next, the side representation vector,  $e$ , is added to the *output layer* before the softmax transformation of the NMT model, e.g.,

$$\begin{aligned} \mathbf{y}_{t+1} &\sim \text{softmax} \left( \mathbf{W}_o \cdot f_t^{dec}(\dots) + \mathbf{b}_e + \mathbf{b}_o \right) \\ \mathbf{b}_e &= \mathbf{W}_e \cdot e; \end{aligned} \quad (3)$$

where  $\mathbf{W}_e \in \mathcal{R}^{|V_T| \times H_s}$  is an additional weight matrix (learnable model parameters) for linear projection of side information representation onto the target output space ( $H_s$  is a predefined dimension for embedding side information). The rationale behind this method is to let the model learn to control the importance of the existing side information contributed to the generation. The function  $f_t^{dec}(\dots)$  is specific to our chosen network reparameterisation, based on RNN (Sutskever et al., 2014; Bahdanau et al., 2015; Luong et al., 2015), or convolution (Gehring et al., 2017b,a; Elbayad et al., 2018), or transformer (Vaswani et al., 2017). Although we make an effort for modification of the NMT model, we believe that it is minimally simple, and generic to suit many different styles of NMT model.

**Multi-task Learning** Consider the case where we would like to use existing side information to

improve the main NMT task. We can define a generative model  $p(y, e|x)$ , formulated as:

$$p(y, e|x) := \underbrace{p(y|x, e)}_{\text{translation model}} \cdot \underbrace{p(e|x)}_{\text{classification model}}; \quad (4)$$

where  $p(y|x, e)$  is a translation model conditioned on the side information as explained earlier;  $p(e|x)$  can be regarded as a classification model – which predicts the side information given the source sentence. Note that side information can often be represented as individual words – which can be treated as labels, making the classification model feasible.

Importantly, the above formulation of a generative model would require summing over “ $e$ ” at test/decode time, which might be done by decoding for all possible label combinations, then reporting the sentence with the highest model score. This may be computationally infeasible in practice. We resort this by approximating the NMT model as  $p(y|x, e) \approx p(y|x)$ , resulting in

$$p(y, e|x) \approx p(y|x) \cdot p(e|x); \quad (5)$$

and thus force the model to encode the shared information in the encoder states.

Our formulation in Equation 5 gives rise to multi-task learning (MTL). Here, we propose the joint learning of two different but related tasks: NMT and multi-label classification (MLC). Here, the MLC task refers to predicting the labels that possibly represent words of the given side information. This is interesting in the sense that the model is capable of not only generating the translated outputs, but also explicitly predicting what the side information is. Here, we adopt a simple instance of MTL for our case, called soft parameter sharing similar to (Duong et al., 2015; Yang and Hospedales, 2016). In our MTL version, the NMT and MLC tasks share the parameters of the encoders. The difference between the two is at the decoder part. In the NMT task, the decoder is kept unchanged. For the MLC task, we define its objective function (or loss), formulated as:

$$\mathcal{L}_{MLC} := - \sum_{m=1}^M \mathbb{1}_{w_m^s}^T \log p_s; \quad (6)$$

where  $p_s$  is the probability of predicting the presence or absence of each element in the side infor-

mation, formulated as:

$$p_s = \text{sigmoid} \left( \mathbf{W}_s \left[ \frac{1}{|\mathbf{x}|} \sum_i g'(x_i) \right] + \mathbf{b}_s \right); \quad (7)$$

where  $\mathbf{x}$  is the source sequence, comprising of  $x_1, \dots, x_i, \dots, x_{|\mathbf{x}|}$  words. Here, we denote a generic function term  $g'(\cdot)$  which refers to a vectorised representation of a specific word depending on designing the network architecture, e.g., stacked bidirectional (forward and backward) networks over the source sequence (Bahdanau et al., 2015; Luong et al., 2015); or a convolutional encoder (Gehring et al., 2017b,a) or a transformer encoder (Vaswani et al., 2017).<sup>7</sup> Further,  $\mathbf{W}_s \in \mathcal{R}^{|V_s| \times H_x}$  and  $\mathbf{b}_s \in \mathcal{R}^{|V_s|}$  are two additional model parameters for linear transformation of the source sequence representation (where  $H_x$  is a dimension of the output of the  $g'(\cdot)$  function, it will differ from network architectures as discussed earlier).

Now, we have two objective functions at the *training* stage, including the NMT loss  $\mathcal{L}_{NMT}$  and the MLC loss  $\mathcal{L}_{MLC}$ . The total objective function of our joint learning will be:

$$\mathcal{L} := \mathcal{L}_{NMT} + \lambda \mathcal{L}_{MLC}; \quad (8)$$

where:  $\lambda$  is the coefficient balancing the two task objectives, whose value is fine-tuned based on the development data to optimise for NMT accuracy measured using BLEU (Papineni et al., 2002).

The idea of MTL applied for NLP was firstly explored by (Collobert and Weston, 2008), later attracts increasing attentions from the NLP community (Ruder, 2017). Specifically, the idea behind MTL is to leverage related tasks which can be learned jointly — potentially introducing an inductive bias (Feinman and Lake, 2018). An alternative explanation of the benefits of MTL is that joint training with multiple tasks acts as an additional regulariser to the model, reducing the risk of overfitting (Collobert and Weston, 2008; Ruder, 2017, *inter alia*).

## 4 Experiments

### 4.1 Datasets

As discussed earlier, we conducted our experiments using three different datasets including TED Talks (Chen et al., 2016), Personalised Europarl

---

<sup>7</sup>Here, to avoid repeating the materials, we will not elaborate their formulations.

	No. of labels	Examples
TED Talks	11	tech business arts issues education health env recreation politics others
Personalised Europarl	2	male female
PatTR-1 (deep)	651	G01G G01L G01N A47F F25D C01B ...
PatTR-2 (shallow)	8	G A F C H B E D

Table 1 Side information statistics for the three datasets, showing the number of types of the side information label, and the set of tokens (display truncated for PatTR-1 (deep)).

(Rabinovich et al., 2017), and PatTR (Wäschle and Riezler, 2012b; Simianer and Riezler, 2013), translating from German (de) to English (en). The statistics of the training and evaluation sets can be shown in Table 2. For the TED Talks and Personalised Europarl datasets, we followed the same sizes of data splits since they are made available on the authors’ github repository and website. For the PatTR dataset, we use the *Abstract* sections for patents from 2008 or later, and the development and test sets are constructed to have 2000 sentences each, similar to (Wäschle and Riezler, 2012b; Simianer and Riezler, 2013).

It is important to note the labeling information for side information. We extracted all kinds of side information from three aforementioned datasets in the form of individual words or labels. This makes the label embeddings much easier. Their relevant statistics and examples can be found in Table 1.

We preprocessed all the data using Moses’s training scripts<sup>8</sup> with standard steps: punctuation normalisation, tokenisation, truecasing. For training sets, we set word-based length thresholds for filtering long sentences since they will not be useful when training the seq2seq models as suggested in the NMT literature (Sutskever et al., 2014; Bahdanau et al., 2015; Luong et al., 2015, *inter alia*). We chose 80, 80, 150 length thresholds for TED Talks, Personalized Europarl, and PatTR datasets, respectively. Note that the 150 threshold indicates that the sentences in the PatTR dataset is in average much longer than in the others. For better handling the OOV problem, we segmented all the preprocessed data with subword units using byte-pair-encoding (BPE) method proposed by Sennrich et al. (2016b). We already know that languages such English and German share an alphabet (Sennrich et al., 2016b), hence learning BPE on the concatenation of source and target

<sup>8</sup><https://github.com/moses-smt/mosesdecoder/tree/master/scripts>

languages (hence called shared BPE) increases the consistency of the segmentation. We applied 32000 operations for learning the shared BPE by using the open-source toolkit.<sup>9</sup> Also, we used dev sets for tuning model parameters and early stopping of the NMT models based on the perplexity. Table 2 shows the resulting vocabulary sizes after subword segmentation for all datasets.

## 4.2 Baselines and Setups

Recall that our method for incorporating the additional side information into the NMT models is generic; hence, it is applicable to any NMT architecture. We chose the transformer architecture (Vaswani et al., 2017) for all our experiments since it arguably is currently the most robust NMT models compared to RNN and convolution based architectures. We re-implemented the transformer - based NMT system using the C++ Neural Network Library - DyNet<sup>10</sup> as our deep learning backend toolkit. Our re-implementation results in the open source toolkit.<sup>11</sup>

In our experiments, we use the same configurations for all transformer models and datasets, including: 2 encoder and decoder layers; 512 input embedding and hidden layer dimensions; sinusoid positional encoding; dropout with 0.1 probability for source and target embeddings, sub-layers (attention + feedforward), attentive dropout; and label smoothing with weight 0.1. For training our neural models, we used early stopping based on development perplexity, which usually occurs after 20-30 epochs.<sup>12</sup>

We conducted our experiments with various incorporation methods as discussed in Section 3. We

<sup>9</sup><https://github.com/rsennrich/subword-nmt>

<sup>10</sup><https://github.com/clab/dynet/>

<sup>11</sup><https://github.com/duyvuleo/Transformer-DyNet/>

<sup>12</sup>The training process of transformer models is much faster than the RNN and convolution - based ones, but requires more epochs for convergence.

dataset	# tokens (M)		# types (K)	# sents	# length limit
<b>TED Talks de→en</b>					
train	3.73	3.75	19.78	14.23	163653
dev	0.02	0.02	4.03	3.15	567
test	0.03	0.03	6.07	4.68	1100
<b>Personalised Europarl de→en</b>					
train	8.46	8.39	21.15	14.04	278629
dev	0.16	0.16	14.67	9.83	5000
test	0.16	0.16	14.76	9.88	5000
<b>PatTR de→en</b>					
train	33.07	32.52	24.97	13.28	656352
dev	0.13	0.13	13.50	6.88	2000
test	0.13	0.12	13.35	6.89	2000

Table 2 Statistics of the training & evaluation sets from datasets including TED Talks, Personalised Europarl, and PatTR; showing in each cell the count for the source language (left) and target language (right); “#types” refers to subword-segmented vocabulary sizes; “n.a.” is not applicable, for development and test sets. Note that all the “#tokens” and “#types” are approximated.

Method	TED Talks	Personalised Europarl	PatTR-1	PatTR-2
<i>base</i>	29.48	31.12	45.86	
<i>si-src-prefix</i>	29.28	30.87	<b>45.99</b>	<b>45.97</b>
<i>si-src-suffix</i>	29.36	31.03	<b>46.01</b>	45.83
<i>si-trg-prefix-p</i>	29.06	30.89	<b>45.97</b>	45.85
<i>si-trg-prefix-h</i>	29.28	30.93	<b>46.03</b>	<b>45.92</b>
<i>output-layer</i>	<b>29.99†</b>	<b>31.22</b>	<b>46.32†</b>	<b>46.09</b>
w/o side info	<b>29.62</b>	31.10	<b>46.14</b>	<b>45.99</b>
<i>mtl</i>	<b>29.86†</b>	31.12	<b>46.14</b>	<b>46.01</b>

Table 3 Evaluation results with BLEU scores of various incorporation variants against the baseline; **bold**: better than the baseline, †: statistically significantly better than the baseline.

denote the system variants as follows:

*base* refers to the baseline NMT system using the transformer without using any side information.

*si-src-prefix* and *si-src-suffix* refer to the NMT system using the side information as respective prefix or suffix of the source sequence, applied to both training and decoder/inference.

*si-trg-prefix* refers to the NMT system using the side information as prefix of the target sequence. There are two variants, including “*si-trg-prefix-p*” means the side information is generated by the model itself and is then used for decoding/inference; “*si-trg-prefix-h*” means the side information is given at decoding/inference runtime.

*output-layer* refers to the method of incorporating side information in the final output layer.

*mtl* refers to the multi-task learning method.

It’s worth noting that the dimensional value for the *output-layer* method was fine-tuned over

the development set, using the value range of {64, 128, 256, 512}. Similarly, the balancing weight in the *mtl* method is fine-tuned using the value range of {0.001, 0.01, 0.1, 1.0}. For evaluation, we measured the end translation quality with case-sensitive BLEU (Papineni et al., 2002). We averaged 2 runs for each of the method variants.

### 4.3 Results and Analysis

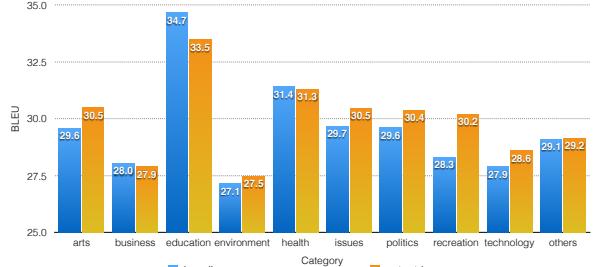
The experimental results can be seen in Table 3. Overall, we obtained limited success for the method of adding side information as prefix or suffix for TED Talks and Personalised Europarl datasets. On the PatTR dataset, small improvements (0.1-0.2 BLEU) are observed. We experimented two sets of side information in the PatTR dataset, including PatTR-1 (651 deep labels) and PatTR (8 shallow labels).<sup>13</sup> The possible reason for this phenomenon is that the multi-head attention mechanism in the transformer may have some confusion given the existing side information, ei-

<sup>13</sup>The shallow setting takes the first character of each label code, which denotes the highest level concept in the type hierarchy, e.g., G01P (measuring speed) → G (physics), with definitions as shown in Fig 3.

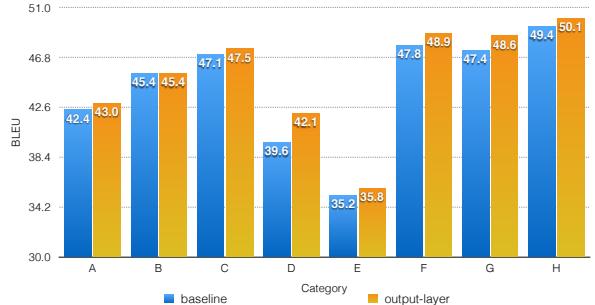
ther in source or target sequences. In some ambiguous cases, the multi-head attention may not know where it should pay more attention. Another possible reason is that the implicit ambiguity of side information that may exist in the data.

Contrary to these variants, the *output-layer* variant was more consistently successful, obtaining the best results across datasets. In the TED Talks and PatTR datasets, this method also provides the statistically significant results compared to the baselines. Additionally, we conducted another experiment by splitting the TED Talks and coarse PatTR-2 datasets by the meta categories, then observed the individual effects when incorporating the side information with *output-layer* variant, as shown in Figure 4a and 4b. In the TED Talks dataset, we observed improvements for most categories, except for “business, education”. In the coarse PatTR-2 dataset, the improvements are obtained across all categories. The key behind this success of the *output-layer* variant is that the representation of existing side information is added in the final output layer and controlled by additional learnable model parameters. In that sense, it results in a more direct effect on lexical choice of the NMT model. This resembles the success in the context of language modelling as presented in Hoang et al. (2016a). Further, we also obtained the promising results for the *mtl* variant although we did implement a very simple instance of MTL with a sharing mechanism and no side information given at a test time. For a fair comparison with the *output-layer* method, we added an additional experiment in which the *output-layer* method does not have the access of side information. As expected, its performance has been dropped, as can be seen in the second last row in Table 3. In this case, the *mtl* method without the side information at a test time performs better. We believe that more careful design of the *mtl* variant can lead to even better results. We also think that the hybrid method combining the *output-layer* and *mtl* variants is also an interesting direction for future research, e.g., relaxing the approximation as shown in Equation 5.

Given the above results, we can find that the characteristics of side information plays an important role in improving the NMT models. Our empirical experiments show that topical information (as in the TED Talks and PatTR datasets) is more useful than the personal traits (as in the Person-



(a) The TEDTalks dataset.



(b) The coarse PatTR-2 dataset.

Fig. 4 Effects on individual BLEU scores for each of categories in the TEDTalks and coarse PatTR-2 datasets, with the NMT model enhanced with the *output-layer* variant.

alised Europarl dataset). However, sometimes it is still good to reserve the personal traits in the target translations (Rabinovich et al., 2017) although their BLEU scores are not necessarily better.

## 5 Related Work

Our work is mainly inspired from Hoang et al. (2016a) who proposed the use of side information for boosting the performance of recurrent neural network language models. We further apply this idea for a downstream task in neural machine translation.

We’ve adapted different methods in the literature for this specific problem and evaluated using different datasets with different kinds of side information.

Our methods for incorporating side information as *suffix*, *prefix* for either source or target sequences have been adapted from (Sennrich et al., 2016a; Johnson et al., 2017).

Also, Michel and Neubig (2018) proposed to personalise neural MT systems by taking the variance that each speaker speaks/writes on his own into consideration. They proposed the adaptation process which takes place in the “output” layer, similar to our *output layer* incorporation method.

The benefit of the proposed MTL approach is not surprising, resembling from existing works, e.g., jointly training translation models from/to multiple languages (Dong et al., 2015); jointly training the encoders (Zoph and Knight, 2016) or both encoders and decoders (Johnson et al., 2017).

## 6 Conclusion

In this work, we have presented various situations to which extent the side information can boost the performance of the NMT models. We have studied different kinds of side information (e.g. topic information, personal trait) as well as present different ways of incorporating them into the existing NMT models. Though being simple, the idea of utilising the side information for NMT is indeed feasible and we have proved it via our empirical experiments. Our findings will encourage practitioners to pay more attention to the side information if exists. Such side information can provide valuable external knowledge that compensates for the learning models. Further, we believe that this idea is not limited to the context of neural LM or NMT, but it may be applicable to other NLP tasks such as summarisation, parsing, reading comprehension, and so on.

## Acknowledgments

We thank the reviewers for valuable feedbacks and discussions. Cong Duy Vu Hoang is supported by Australian Government Research Training Program Scholarships at the University of Melbourne, Australia.

## References

- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *Proc. of 3rd International Conference on Learning Representations (ICLR2015)*.
- Iacer Calixto and Qun Liu. 2017. Incorporating global visual features into attention-based neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*. pages 992–1003. <https://aclanthology.info/papers/D17-1105/d17-1105>.
- Iacer Calixto, Qun Liu, and Nick Campbell. 2017. Doubly-attentive decoder for multi-modal neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*. pages 1913–1924. <https://doi.org/10.18653/v1/P17-1175>.
- Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. WIT<sup>3</sup>: Web Inventory of Transcribed and Translated Talks. In *Proceedings of the 16<sup>th</sup> Conference of the European Association for Machine Translation (EAMT)*. Trento, Italy, pages 261–268.
- Wenhu Chen, Evgeny Matusov, Shahram Khadivi, and Jan-Thorsten Peter. 2016. Guided alignment training for topic-aware neural machine translation. *CoRR* abs/1607.01628. <http://arxiv.org/abs/1607.01628>.
- Trevor Cohn, Cong Duy Vu Hoang, Ekaterina Vymolova, Kaisheng Yao, Chris Dyer, and Gholamreza Haffari. 2016. Incorporating Structural Alignment Biases into an Attentional Neural Translation Model. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, San Diego, California, pages 876–885. <http://www.aclweb.org/anthology/N16-1102>.
- Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning*. ACM, New York, NY, USA, ICML ’08, pages 160–167. <https://doi.org/10.1145/1390156.1390177>.
- Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In *ACL (1)*. The Association for Computer Linguistics, pages 1723–1732.
- Long Duong, Trevor Cohn, Steven Bird, and Paul Cook. 2015. Low resource dependency parsing: Cross-lingual parameter sharing in a neural network parser. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Association for Computational Linguistics, pages 845–850. <https://doi.org/10.3115/v1/P15-2139>.
- Maha Elbayad, Laurent Besacier, and Jakob Verbeek. 2018. Pervasive attention: 2d convolutional neural networks for sequence-to-sequence prediction. *CoRR* abs/1808.03867. <http://arxiv.org/abs/1808.03867>.
- Reuben Feinman and Brenden M. Lake. 2018. Learning inductive biases with simple neural networks. *CoRR* abs/1802.02745. <http://arxiv.org/abs/1802.02745>.
- Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. Language, Speech, and Communication. MIT Press, Cambridge, MA.

- Mercedes García-Martínez, Loïc Barrault, and Fethi Bougares. 2017. Neural machine translation by generating multiple linguistic factors. *CoRR* abs/1712.01821. <http://arxiv.org/abs/1712.01821>.
- Jonas Gehring, Michael Auli, David Grangier, and Yann Dauphin. 2017a. A convolutional encoder model for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*. pages 123–135. <https://doi.org/10.18653/v1/P17-1012>.
- Jonas Gehring, Michael Auli, David Grangier, Dennis Yarats, and Yann N. Dauphin. 2017b. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*. pages 1243–1252. <http://proceedings.mlr.press/v70/gehring17a.html>.
- Cong Duy Vu Hoang, Trevor Cohn, and Gholamreza Haffari. 2016a. Incorporating Side Information into Recurrent Neural Network Language Models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, San Diego, California, pages 1250–1255. <http://www.aclweb.org/anthology/N16-1149>.
- Cong Duy Vu Hoang, Reza Haffari, and Trevor Cohn. 2016b. Improving Neural Translation Models with Linguistic Factors. In *Proceedings of the Australasian Language Technology Association Workshop 2016*. Melbourne, Australia, pages 7–14. <http://www.aclweb.org/anthology/U16-1001>.
- Melvin Johnson, Mike Schuster, Quoc Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viegas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google’s multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics* 5:339–351. <https://transacl.org/ojs/index.php/tacl/article/view/1081>.
- Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective Approaches to Attention-based Neural Machine Translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, pages 1412–1421. <http://aclweb.org/anthology/D15-1166>.
- Evgeny Matusov, Andy Way, Iacer Calixto, Daniel Stein, Pintu Lohar, and Sheila Castilho. 2017. Using images to improve machine-translating e-commerce product listings. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 2: Short Papers*. pages 637–643. <https://aclanthology.info/papers/E17-2101/e17-2101>.
- Haitao Mi, Baskaran Sankaran, Zhiguo Wang, and Abe Ittycheriah. 2016. Coverage Embedding Models for Neural Machine Translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, pages 955–960.
- Paul Michel and Graham Neubig. 2018. Extreme adaptation for personalized neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, pages 312–318. <http://aclweb.org/anthology/P18-2050>.
- Shachar Mirkin, Scott Nowson, Caroline Brun, and Julien Perez. 2015. Motivating personality-aware machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 1102–1108. <https://doi.org/10.18653/v1/D15-1130>.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, Stroudsburg, PA, USA, ACL ’02, pages 311–318. <https://doi.org/10.3115/1073083.1073135>.
- Ella Rabinovich, Raj Nath Patel, Shachar Mirkin, Lucia Specia, and Shuly Wintner. 2017. Personalized machine translation: Preserving original author traits. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Association for Computational Linguistics, pages 1074–1084. <http://aclweb.org/anthology/E17-1101>.
- Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *CoRR* abs/1706.05098. <http://arxiv.org/abs/1706.05098>.
- Rico Sennrich and Barry Haddow. 2016. Linguistic input features improve neural machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 1, Research Papers*. Association for Computational Linguistics, pages 83–91. <https://doi.org/10.18653/v1/W16-2209>.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Controlling politeness in neural machine translation via side constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, pages 35–40. <https://doi.org/10.18653/v1/N16-1005>.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume*

*1: Long Papers*). Association for Computational Linguistics, Berlin, Germany, pages 1715–1725. <http://www.aclweb.org/anthology/P16-1162>.

Patrick Simianer and Stefan Riezler. 2013. Multi-task learning for improved discriminative training in SMT. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Sofia, Bulgaria, pages 292–300. <http://www.aclweb.org/anthology/W13-2236>.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to Sequence Learning with Neural Networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*. MIT Press, Cambridge, MA, USA, NIPS’14, pages 3104–3112.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, Curran Associates, Inc., pages 5998–6008. <http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf>.

Katharina Wäschle and Stefan Riezler. 2012a. Analyzing Parallelism and Domain Similarities in the MAREC Patent Corpus. *Multidisciplinary Information Retrieval* pages 12–27. <http://www.cl.uni-heidelberg.de/~riezler/publications/papers/IRF2012.pdf>.

Katharina Wäschle and Stefan Riezler. 2012b. Structural and Topical Dimensions in Multi-Task Patent Translation. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012)*, Avignon, France. <http://www.aclweb.org/anthology-new/E/E12/E12-1083.pdf>.

Yongxin Yang and Timothy M. Hospedales. 2016. Trace norm regularised deep multi-task learning. *CoRR* abs/1606.04038. <http://arxiv.org/abs/1606.04038>.

Jiacheng Zhang, Yang Liu, Huanbo Luan, Jingfang Xu, and Maosong Sun. 2017. Prior knowledge integration for neural machine translation using posterior regularization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pages 1514–1523. <https://doi.org/10.18653/v1/P17-1139>.

B. Zoph and K. Knight. 2016. Multi-source neural translation. In *Proc. NAACL*. <http://www.isi.edu/natural-language/mt/multi-source-neural.pdf>.

# Text-dependent Forensic Voice Comparison: Likelihood Ratio Estimation with the Hidden Markov Model (HMM) and Gaussian Mixture Model – Universal Background Model (GMM-UBM) Approaches

Satoru Tsuge

Daido University, Japan

tsuge@daido-it.ac.jp

Shunichi Ishihara

Australian National University

shunichi.ishihara@anu.edu.au

## Abstract

Among the more typical forensic voice comparison (FVC) approaches, the acoustic-phonetic statistical approach is suitable for text-dependent FVC, but it does not fully exploit available time-varying information of speech in its modelling. The automatic approach, on the other hand, essentially deals with text-independent cases, which means temporal information is not explicitly incorporated in the modelling. Text-dependent likelihood ratio (LR)-based FVC studies, in particular those that adopt the automatic approach, are few. This preliminary LR-based FVC study compares two statistical models, the Hidden Markov Model (HMM) and the Gaussian Mixture Model (GMM), for the calculation of forensic LRs using the same speech data. FVC experiments were carried out using different lengths of Japanese short words under a forensically realistic, but challenging condition: only two speech tokens for model training and LR estimation. Log-likelihood-ratio cost ( $C_{llr}$ ) was used as the assessment metric. The study demonstrates that the HMM system constantly outperforms the GMM system in terms of average  $C_{llr}$  values. However, words longer than three mora are needed if the advantage of the HMM is to become evident. With a seven-mora word, for example, the HMM outperformed the GMM by a  $C_{llr}$  value of 0.073.

## 1 Introduction

After the DNA success story, the likelihood ratio (LR)-based approach became the new paradigm for evaluating and presenting forensic evidence in court. The LR approach has also been applied to speech evidence (Rose, 2006), and it is increasing-

ly accepted in forensic voice comparison (FVC) as well (Morrison, 2009).

There are two different approaches in FVC. They are the ‘acoustic-phonetic statistical approach’ and the ‘automatic approach’ (Morrison et al., 2018). The former usually works on comparable phonetic units that can be found in both the offender and suspect samples. In the latter, acoustic measurements are usually carried out over all portions of the available recordings, resulting in more detailed acoustic characteristics of the speakers. The common statistical models used in the automatic approach are the Gaussian mixture model – universal background model (GMM-UBM) (Reynolds et al., 2000) and i-vectors with probabilistic linear discrimination analysis (PLDA) (Burget et al., 2011). Due to its nature, the automatic approach is mainly used for text-‘independent’ FVC, and there is a good amount of research on this (Enzinger & Morrison, 2017; Enzinger et al., 2016). The acoustic-phonetic statistical approach is a type of text-‘dependent’ FVC because it tends to focus on particular linguistic units, such as phonemes, words, phrases, etc. Having said that, even if one is targeting a particular word or phrase, for example ‘hello’, all obtainable features are not exploited in the acoustic-phonetic statistical approach because it still tends to focus on particular segments or phonemes of the word or phrase, e.g. the formant trajectories of the diphthong and the static spectral information of the fricative (Rose, 2017).

One of the advantages of text-dependent FVC is the availability of the time-varying characteristics of a speaker, which is information that can be explicitly included in the modelling.

There are a good number of LR-based text-independent FVC studies in the automatic approach (Enzinger & Morrison, 2017; Enzinger et al., 2016). However, although there are some stud-

ies in which text-independent models (e.g. GMM) were applied to text-dependent FVC scenarios (Morrison, 2011), to the best of our knowledge, studies on LR-based text-dependent FVC in the automatic approach are scarce.

In this study, a text-dependent LR-based FVC system with the GMM-UBM based system (GMM system) and that with the hidden Markov model (HMM system) are compared in their performance using the same data. The transitional characteristic of individual speech can be explicitly modelled in the latter system.

Words of various length are used for testing purposes to see how word duration influences the performance of the systems. Having the forensically realistic condition of data sparsity in mind, we used only two tokens of each word for modelling and testing.

It is naturally expected that, given a sufficient amount of data, the HMM system outperforms the GMM system. However, it is not so clear whether the above expectation is realistic when the amount of data is limited. Even if the HMM system works better, it is important to establish how the HMM and GMM systems compare with respect to the calculation of strength of LR, and also how and under what conditions the former is more advantageous than the latter.

## 2 Likelihood Ratios

The LR framework has been advocated by many as the logically and legally correct framework for assessing forensic evidence and reporting the outcome in court (Aitken, 1995; Aitken & Stoney, 1991; Aitken & Taroni, 2004; Balding & Steele, 2015; Evett, 1998; Robertson & Vignaux, 1995). A substantial amount of fundamental research on FVC has been carried out since the late 1990s (Gonzalez-Rodriguez et al., 2007; Morrison, 2009; Rose, 2006), and it is now accepted in an increasing number of countries (Morrison et al., 2016).

In the LR framework, the task of the forensic expert is to estimate strength of evidence and report it to the court. LR is a measure of the quantitative strength of evidence, and is calculated using the formula in 1).

$$LR = \frac{p(E|H_p)}{p(E|H_d)} \quad 1)$$

In 1), E is the evidence, i.e. the measured properties of the voice evidence;  $p(E|H_p)$  is the proba-

bility of E, given  $H_p$ , in other words the prosecution or same-speaker hypothesis;  $p(E|H_d)$  is the probability of E, given  $H_d$ , in other words the defence or different-speaker hypothesis (Robertson & Vignaux, 1995). The LR can be considered in terms of the ratio between similarity and typicality. Similarity here means the similarity of evidence attributable to the offender and the suspect, respectively. Typicality means the typicality of that evidence against the relevant population.

The relative strength of the given evidence with respect to the competing hypotheses ( $H_p$  vs.  $H_d$ ) is reflected in the magnitude of the LR. If the evidence is more likely to occur under the prosecution hypothesis than under the defence hypothesis, the LR will be higher than 1. If the evidence is more likely to occur under the defence hypothesis than under the persecution hypothesis, the LR will be lower than 1. For example,  $LR = 30$  means that the evidence is 30 times more likely to occur on the assumption that the evidence is from the same person than on the assumption that it is not.

The important point is that the LR is concerned with the probability of the evidence, given the hypothesis (either  $H_p$  or  $H_d$ ). The probability of the evidence can be estimated by forensic scientists. They legally must not and logically cannot estimate the probability of the hypothesis, given the evidence. This is because the forensic scientist is not legally in a position to refer to the ultimate ‘guilty vs. non-guilty’ question, i.e. the probability of the hypothesis, given the evidence. That is the task of the trier-of-fact. Furthermore, the forensic scientist would need to refer to the Bayesian theorem to estimate the probability of the hypothesis, given the evidence, using prior information that is only accessible to the trier-of-fact; thus the forensic scientist cannot logically estimate the probability of the hypothesis.

## 3 Experimental Design

In this section, the nature of the database used for the experiments is explained first. This is followed by an illustration as to how the speaker comparisons were set up for the experiments. The acoustic features used in this study will be explained towards the end.

### 3.1 Database

Our data were extracted from the National Research Institute of Police Science (NRIPS) data-

ze.ro 'zero'	ku.ru.ma 'car'	ko.o.so.ku 'highway'	hya.ku 'hundred'	ka.ne 'money';=	go.ze.n 'AM'
re.e 'zero'	de.n.wa 'telephone'	ya.ku.so.ku 'promise'	sa.n.bya.ku 'three hundred'	da.i.jyo.o.bu 'fine'	wa.ta.shi 'I'
i.chi 'one'	ke.e.sa.tsu 'police'	o.n.na 'woman'	ro.p.pya.ku 'six hundred'	ki.no.o 'yesterday'	ko.do.mo 'child'
sa.n 'three'	do.ku 'poison'	o.ku.sa.n 'wife'	ha.p.pya.ku 'eight hundred'	kyo.o 'today'	ke.e.ta.i 'mobile phone'
yo.n 'four'	re.n.ra.ku 'contact'	re.su.to.ra.n 'restaurant'	se.n 'thousand'	a.shi.ta 'tomorrow'	ka.ji 'fire'
ro.ku 'six'	ba.ku.da.n 'bomb'	po.su.to 'post'	i.s.se.n 'one thousand'	ge.n.ki.n 'cash'	ko.n.bi.ni 'store'
na.na 'seven'	gi.n.ko.o 'bank'	sa.a.bi.su.e.ri.a 'road house'	go.go 'afternoon'	a.no.o 'well (filler)'	ta.ku.shi.i 'taxi'
shi.chi 'seven'	ji.ka.n 'time'	sa.n.z.e.n 'three thousand'	e.ki 'station'	ne.e 'well (filler)'	i.n.ta.a 'interchange'
ha.chi 'eight'	mo.shi.mo.shi 'hello (phone)'	ha.s.se.n 'eight thousand'	o.ma.e 'you'	a.no.ne.e 'well (filler)''	me.e.ru 'mail'
kyu.u 'nine'	ha.i 'yes'	ma.n 'ten thousand'	o.i 'hay'	na.ka.ma 'mate'	ba.n.go.o 'number'
jyu.u 'ten'	o.re. 'I'	o.ku 'million'	ba.ku.ha.tsu 'explosion'	ka.i.sha 'company'	ko.o.za 'account'

Table 1: 66 target words with their glosses. Each mora is separated by a period.

base (Makinae et al., 2007). The database consists of recordings collected from 316 male and 323 female speakers. All utterances were read-out speech, consisting of single syllables, words, selected sentences and so on. The word-based recordings stored in the database provided the data used in this study.

Participants ranged in age from 18 to 76 years. The metadata provide information on the areas of Japan (or overseas in some cases) where they have resided, as well as their height, weight, and their health conditions on the day of recording. Only male speakers who completed the recordings in two different sessions separated by 2-3 months, without any mis-recordings for the target 66 words, were selected for the current study (resulting in 310 speakers). Each word was recorded only twice in each session.

The rhythmic unit of Japanese is the mora. Based on mora, the 66 words, all listed in Table 1, consist of 25 two-, 16 three-, 22 four-, 2 five- and 1 seven-mora words.

The 310 speakers were separated into six different, mutually exclusive groups: Gr1 (59 speakers), Gr2 (60), Gr3 (60), Gr4 (60), Gr5 (60) and Gr6 (13). Five different experiments were conducted using the six groups, as shown in Table 2.

The test database was used for simulating two types of offender-suspect comparisons: same-speaker (SS) and different-speaker (DS). An LR

was estimated for each of the comparisons. The development database was also called upon for simulating offender-suspect comparisons, but the derived scores (pre-calibration LRs) were specifically used to obtain the weights for calibration (refer to §4.4 for details on calibration). The background database was used to build the statistical model for typicality.

Experiments	Test	Dev	Back
Exp1	Gr1	Gr2	Gr3,4,5,6
Exp2	Gr2	Gr3	Gr1,4,5,6
Exp3	Gr3	Gr4	Gr1,2,5,6
Exp4	Gr4	Gr5	Gr1,2,3,6
Exp5	Gr5	Gr1	Gr2,3,4,6

Table 2: Usage of Gr1~6 for experiments (Exp). Test, Dev and Back refer to test, development and background databases.

As mentioned earlier, there are two recordings per speaker for each word in each session. The suspect model was built using two recordings taken from one session, and an LR was estimated for each of the two recordings of the other session (offender evidence). The same process was repeated by swapping the recordings of the sessions. In this way, 4 LRs were obtained for each SS comparison, and 8 LRs for each DS comparison. Thus, the number of comparisons is  $4^*n$  ( $n$  = number of speakers) for the SS comparisons, and  $8^*nC_2$  ( $C$ =combination) for the DS comparisons. Using the five different groups (Gr1~5) separately

as a test database, it was possible, altogether, to carry out 1188 SS comparisons and 69392 DS comparisons. The breakdowns of the SS and DS comparisons are given in Table 3 for the five experiments (Exp1~5).

Experiments	SS	DS
Exp1: Gr1 (59)	236	13688
Exp2: Gr2 (60)	240	14160
Exp3: Gr3 (60)	240	14160
Exp4: Gr4 (58)	232	13224
Exp5: Gr5 (60)	240	14160
<b>Total</b>	<b>1188</b>	<b>69392</b>

Table 3: Numbers of SS and DS comparisons for each word.

The NRIPS database also contains the recordings of 50 sentences that are based on ATR phonetically balanced Japanese sentences (Kurematsu et al., 1990). These sentences were used to build the initial statistical models (refer to §4.1 and §4.2 for details).

### 3.2 Acoustic Features

Twelve mel-frequency cepstral coefficients (MFCCs), 12  $\Delta$  MFCCs and  $\Delta$  log power (a feature vector of 25<sup>th</sup>-order) were extracted with a 20 msec wide hamming window shifting every 10 msec.

## 4 Estimation of Likelihood Ratios

In this section, the two different modelling techniques used in the current study are explained. This is followed by an exposition of the method for calculating scores with these models. The method used for converting the scores to the LRs, namely calibration, will be explained last.

For this study, the suspect model, rather than being based solely on the data of the suspect speaker, was generated by adapting a speaker-unspecific model (background model) by means of a maximum a posteriori (MAP) procedure. Three different numbers of Gaussians (4, 8 and 16) were tried in the models.

### 4.1 GMM Models

The following is the process of building a speaker-specific word-dependent GMM for each speaker.

- 1) To build a speaker-unspecific word-independent GMM using the recordings of the phonetically balanced utterances;
- 2) To build a speaker-unspecific word-dependent GMM for each word by training the speaker-

unspecific word-independent GMM, which was generated in 1), with the relevant word recordings of the background database;

- 3) To build the speaker-specific word-dependent GMM (suspect model =  $\lambda_{sus}$ ) for each word by training the speaker-unspecific word-independent GMM, which was built in 2), with the speaker specific data in the test database, while applying a MAP adaptation.

The speaker-unspecific word-dependent GMM, which was built in 2) for each word, was used as the background model ( $\lambda_{bkg}$ ).

### 4.2 HMM Models

The following is the process of building a speaker-specific word-dependent HMM for each speaker.

- 1) To build speaker-unspecific phoneme-dependent HMMs using the recordings of the phonetically balanced utterances;
- 2) To build an initial speaker-unspecific word-dependent HMM for each word by concatenating speaker-unspecific phoneme-dependent HMMs, which were built in 1).
- 3) To build speaker-specific word-dependent HMM (suspect model =  $\lambda_{sus}$ ) by training the initial speaker-unspecific word-dependent HMM, which was built in 2), with the speaker specific data in the test database, while applying a MAP adaptation.

The initial speaker-unspecific word-dependent HMM, which was built in 2), was trained with the relevant word recordings of the background database, and the resultant model was used as the speaker-unspecific word-dependent background model ( $\lambda_{bkg}$ ).

### 4.3 Score Calculations

The score of each comparison can be estimated using the equation given in 2), in which  $s$  = score,  $x_t$  = an observation sequence of vectors of acoustic features constituting the offender data of which there are a total of  $T$ ,  $\lambda_{sus}$  = suspect model and  $\lambda_{bkg}$  = background model.

$$s = \frac{1}{T} \sum_{t=1}^T \log(p(x_t | \lambda_{sus})) - \log(p(x_t | \lambda_{bkg})) \quad 2)$$

A score is estimated as the mean of the relative values of the two probability density functions for the feature vectors extracted from the offender data, and was calculated for each of the SS and DS comparisons.

#### 4.4 Scores to Likelihood Ratios

The outcomes of the GMM and HMM systems are not LRs, but are known as *scores*. The value of a score provides information about the degree of the similarity between the two speech samples, i.e. the offender and suspect samples, having taken into account their typicality with respect to the relevant population; it is not directly interpretable as an LR (Morrison, 2013, p. 2). Thus, the scores need to be converted to LRs by means of a calibration process. As we will see in §6, calibration is an essential part of LR-based FVC.

Logistic-regression calibration (Brümmer & du Preez, 2006) is a commonly used method that converts scores to interpretable LRs by applying linear shifting and scaling in the log odds space. A logistic-regression line (e.g.  $y = ax + b$ ;  $x$  = score;  $y = \log_{10}LR$ ) whose weights (i.e.  $a$  and  $b$  in  $y = ax + b$ ) are estimated from the SS and DS scores of the development database is used to monotonically shift (by the amount of  $b$ ) and scale (by the amount of  $a$ ) the scores of the testing database to the  $\log_{10}$ LRs.

### 5 Assessment Metrics

A common way of assessing the performance of a classification system is with reference to its correct- or incorrect-classification rate: for instance, how many of the SS comparisons were correctly assessed as coming from the same speakers, and how many of the DS comparisons were correctly assessed as coming from different speakers. In the context of LR-based FVC, an LR can be used as a classification function with  $LR = 1$  as unity. However, correct- or incorrect-classification rate is a binary decision (same speaker or different speakers), which refers to the ultimate issue of ‘guilty vs. non-guilty’. As explained in §2, it is not the task of the forensic expert, but of the trier-of-fact, to make such a decision. Thus, any metrics based on binary decision are not coherent with the LR framework.

As emphasised in §2, the task of the forensic expert is to estimate the strength of evidence as accurately as possible, and the strength of evidence, which can be quantified by means of a LR,

is not binary in nature, but continuous. For example, both  $LR = 10$  and  $LR = 20$  support the correct hypothesis for the SS comparisons, but the latter supports the hypothesis more strongly than the former. The relative strength of the LR needs to be taken into account in the assessment.

Hence, in this study, the log-likelihood-ratio cost ( $C_{llr}$ ), which is a gradient metric based on LR, was used as the metric for assessing the performance of the LR-based FVC system. The calculation of  $C_{llr}$  is given in 3) (Brümmer & du Preez, 2006).

$$C_{llr} = \frac{1}{2} \left( \frac{1}{N_{H_p}} \sum_{i \text{ for } H_p=\text{true}}^{N_{H_p}} \log_2 \left( 1 + \frac{1}{LR_i} \right) + \frac{1}{N_{H_d}} \sum_{j \text{ for } H_d=\text{true}}^{N_{H_d}} \log_2 (1 + LR_j) \right) \quad 3)$$

In 3),  $N_{H_p}$  and  $N_{H_d}$  are the number of SS and DS comparisons, and  $LR_i$  and  $LR_j$  are the linear LRs derived from the SS and DS comparisons, respectively. Under a perfect system, all SS comparisons should produce LRs greater than 1, since origins are identical; as, in the case of DS comparisons, origins are different, DS comparisons should produce LRs less than 1.  $C_{llr}$  takes into account the magnitude of derived LR values, and assigns them appropriate penalties. In  $C_{llr}$ , LRs that support the counter-factual hypotheses or, in other words, contrary-to-fact LRs ( $LR < 1$  for SS comparisons and  $LR > 1$  for DS comparisons) are heavily penalised and the magnitude of the penalty is proportional to how much the LRs deviate from unity. Optimum performance is achieved when  $C_{llr} = 0$  and decreases as  $C_{llr}$  approaches and exceeds 1. Thus, the lower the  $C_{llr}$  value, the better the performance.

The  $C_{llr}$  measures the overall performance of a system in terms of validity based on a cost function in which there are two main components of loss: discrimination loss ( $C_{llr}^{min}$ ) and calibration loss ( $C_{llr}^{cal}$ ) (Brümmer & du Preez, 2006). The former is obtained after the application of the so-called pooled-adjacent-violators (PAV) transformation – an optimal non-parametric calibration procedure. The latter is obtained by subtracting the former from the  $C_{llr}$ . In this study, besides  $C_{llr}$ ,  $C_{llr}^{min}$  and  $C_{llr}^{cal}$  are also referred to.

The magnitude of the derived LRs is visually presented using Tippett plots. Details on how to read a Tippett plot are explained in §6, when the plots are presented.

## 6 Experimental Results and Discussions

The average  $C_{llr}$ ,  $C_{llr}^{min}$  and  $C_{llr}^{cal}$  values were calculated according to the mora numbers; they are plotted in Figure 1 as a function of word duration, separately for the HMM and GMM systems. The numerical values of Figure 1 are given in Table 4.

		2	3	4	5	7
$C_{llr}$	G	0.309	0.239	0.182	0.146	0.136
	H	0.302	0.230	0.156	0.114	0.063
$C_{llr}^{min}$	G	0.270	0.206	0.152	0.118	0.099
	H	0.261	0.192	0.126	0.085	0.047
$C_{llr}^{cal}$	G	0.038	0.032	0.029	0.028	0.037
	H	0.040	0.037	0.030	0.028	0.016

Table 4: Numerical information of Figure 1.  
G = GMM and H = HMM.

Although this was expected, it can be seen from Figure 1a and Table 4 that the overall performance ( $C_{llr}$ ) of both systems improves as the words become longer in terms of mora, and also that the HMM system constantly outperforms the GMM system as far as average  $C_{llr}$  values are concerned. The performance gap between the two systems becomes wider as the number of mora increases, with the performance of the two systems being similar with words of two and three moras. For 12 out of the 25 two-mora words and 6 out of the 16 three-mora words, the GMM system performed better than the HMM system in terms of  $C_{llr}$ . In other words, the evidence suggests that the HMM may not be clearly advantageous for short words, e.g. two- or three-mora words. For the sake of reference, for only 6 out of the 22 four-mora words, the GMM system outperformed the HMM system. For the five- and seven-mora words, the HMM system constantly outperformed the GMM system.

The discriminability of the systems ( $C_{llr}^{min}$ ) (Figure 1b) also exhibits the same trend as the overall performance in that discriminability improves with more moras, the HMM system constantly performed better than the GMM system, and the performance of the former improves at a faster rate than that of the latter. As a result, there is a larger gap in discriminability between the two systems with the seven-mora word ( $0.052 = 0.099 - 0.047$ ) than there is with the two-mora words ( $0.009 = 0.270 - 0.261$ ).

The calibration loss of both systems ( $C_{llr}^{cal}$ ) (Figure 1c) is very similar for two-, three-, four- and five-mora words, which are essentially the same for the two systems (2: 0.038 and 0.040; 3: 0.032 and 0.037; 4: 0.029 and 0.030; 5: 0.028 and

0.028). The calibration loss improves (albeit at a very small rate) as a function of word duration, except in the case of the GMM system with the seven-mora word.

As has been described by means of Figure 1 and Table 4, it is clearly advantageous to include temporal information in modelling in Japanese, even under the challenging condition of data sparsity. However, the difference in performance may not be evident with short, e.g. two- and three-mora, words. Put differently, if a forensic speech expert is working on a comparable word or phrase of relatively good length, the decision to either include transitional information in the modelling or not is likely to substantially impact on the outcome. For example, the HMM system outperformed the GMM system by the  $C_{llr}$  values of 0.073 (= 0.136 - 0.063) with the seven-mora word.

Three different numbers of Gaussians – 4, 8 and 16 – were used in the study. Table 5 shows which mixture number of Gaussians performed best for words of different mora duration according to the different systems. For example, out of the 25 two-mora words, the GMM system with a mixture number of 8 ( $M = 8$ ) returned the best result for 11 words, and the HMM system with a mixture number of 4 ( $M = 4$ ) yielded the lowest  $C_{llr}$  value for 13 words.

Mora	System	M = 4	M = 8	M = 16
2 (25)	G	0	11	14
	H	13	5	7
3 (16)	G	0	2	14
	H	13	3	0
4 (22)	G	0	1	21
	H	14	4	3
5 (2)	G	0	2	0
	H	1	1	1
7 (1)	G	0	0	1
	H	0	0	1
Total	G	0 (0%)	16 (24%)	50 (76%)
	H	41 (62%)	13 (20%)	12 (18%)

Table 5: Best-performing Gaussian numbers (M) for words with different mora numbers.  
G = GMM and H = HMM.

According to Table 5, there is a clear difference between the two systems with respect to the best performing mixture number of Gaussians, in that the GMM tends to require a higher mixture number for optimal performance (overall, 76% of words worked best with a mixture number of 16), while the HMM generally does not require a

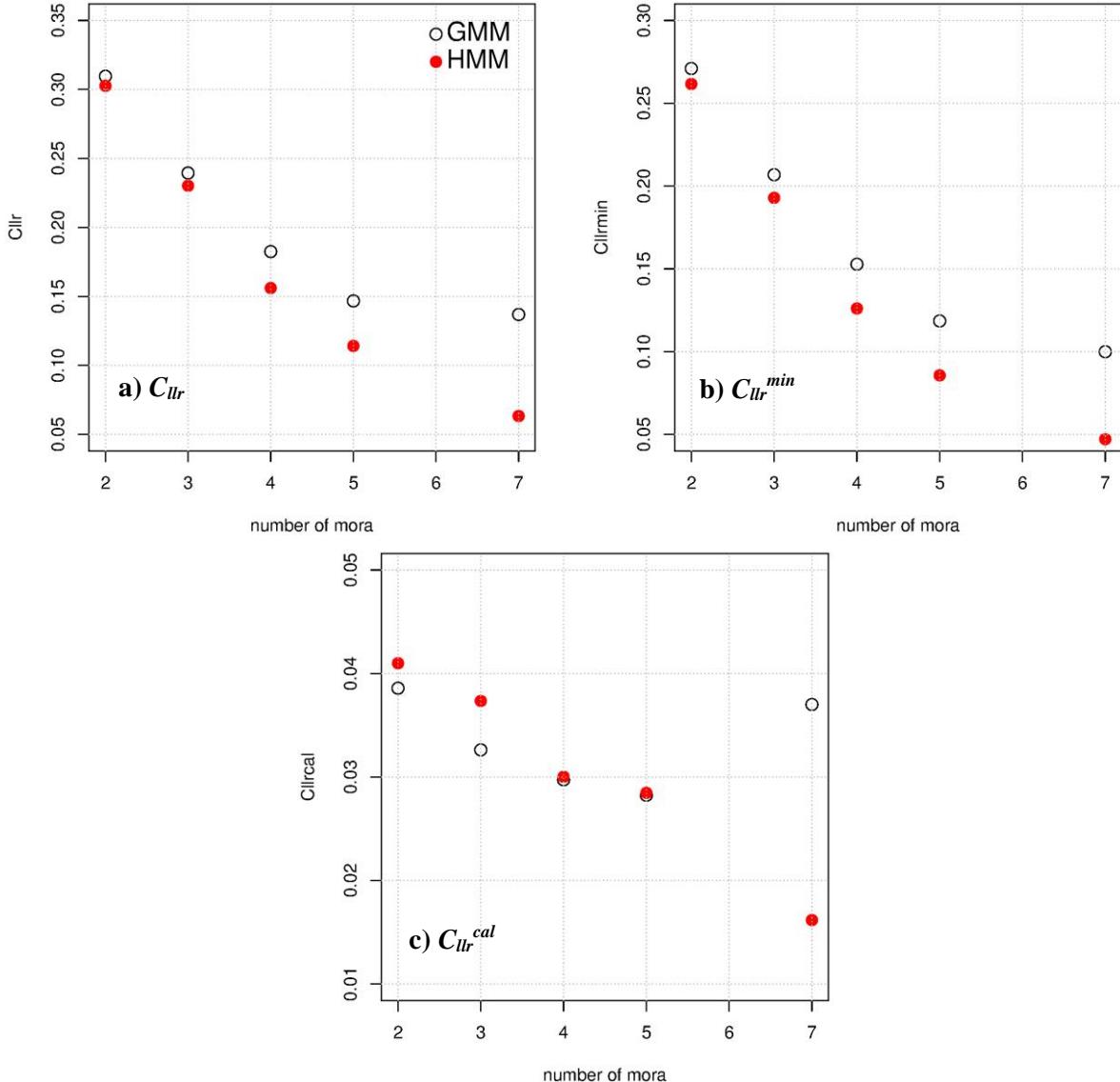


Figure 1:  $C_{llr}$  (Panel a),  $C_{llr}^{min}$  (b) and  $C_{llr}^{cal}$  (c) values are plotted as a function of mora duration, separately for GMM (empty circle) and HMM (filled circle) systems. Note that the Y-axis scale in Panel c is different from that in Panels a and b.

higher mixture number (overall, 62% of words performed best with a mixture number of 4).

To investigate whether there are any differences in the nature and magnitude of the derived LRs, a Tippett plot was generated for each word in each experiment, and this was done separately for the GMM and HMM systems. Figure 2 has Tippett plots of the five-mora word ‘daijyoobu’ with 16 Gaussians: Panel a) = GMM and Panel b) = HMM. The plots are fairly typical and illustrate the differences between the two systems.

Tippett plots show the cumulative proportion of the LRs of the DS comparisons (DSLRs), which are plotted rising from the right, as well as of the LRs of the SS comparisons (SSLRs), plotted rising from the left. The solid curves are for LRs and

the dotted curves are for scores (pre-calibration LRs). For all Tippett plots, the cumulative proportion of trails is plotted on the y-axis against the  $\log_{10}$  LRs on the x-axis.

As can be seen in Figure 2, the derived scores (pre-calibration LRs), which are given in dotted curves, are uncalibrated in different ways for the GMM and HMM systems: the former (Figure 2a) is uncalibrated to the left and the latter (Figure 2b) is uncalibrated to the right. This indicates that calibration is essential in both systems. In fact, calibrating system output is recommended as standard practice (Morrison, 2018).

The dotted curves are more widely apart in Figure 2a (GMM) than in Figure 2b (HMM). This means that the magnitude of the derived scores is

greater with the GMM system than with the HMM system. However, after calibration (solid curves), it can be seen that the magnitude of the DSLRs is very similar between the two systems while the SSLRs are far stronger for the HMM system than for the GMM system. That is, the calibration causes different effects in the two systems; it brings about more conservative LRs for the GMM system, but enhanced LRs for the HMM system.

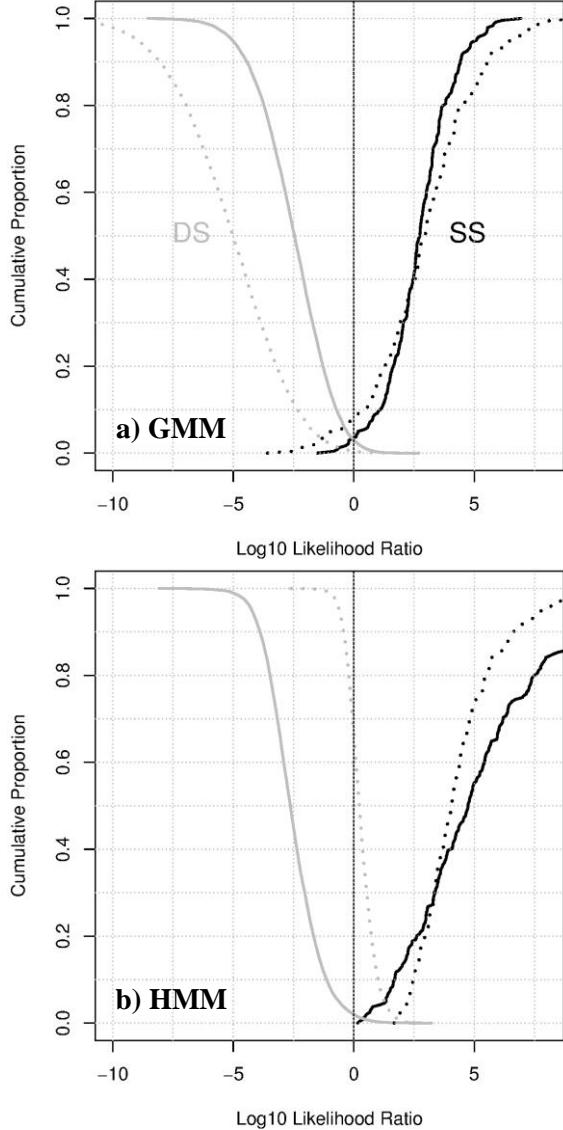


Figure 2: Tippett plots of the five-mora word ‘daijyoobu’ (Exp5) with 16 Gaussians: Panel a) = GMM and Panel b) = HMM

Although calibration usually results in a better performance, its impact on the magnitude of LRs seems to be different depending on various factors, including the types of features and modelling techniques. Many FVC studies, in particular those based on the acoustic-phonetic statistical approach, report that calibration results in more con-

servative LRs than scores (Rose, 2013), while it contributes to stronger LRs for the automatic approach (Morrison, 2018). However, it is not clear at this stage what the observed differences between the two systems with respect to the relationship between the scores and LRs entail. This warrants further investigation.

Apart from the similar degree of magnitude of the DSLRs (including both consistent-with-fact and contrary-to-fact LRs) that were obtained for the GMM and HMM systems, Figure 2 shows that the magnitude of the consistent-with-fact SSLRs is far greater for the HMM system (Figure 2b), and also that all of the SS comparisons were accurately classified as being from the same speakers for the HMM system. As a result, the HMM system is assessed to be better in  $C_{llr}$  than the GMM system (GMM:  $C_{llr} = 0.182$  and HMM:  $C_{llr} = 0.156$ ).

## 7 Conclusions

This is a preliminary study investigating the usefulness of speaker-individuating information manifested in the time-varying aspect of speech in a text-dependent FVC system, in particular in the automatic FVC approach. In this study, performance of the GMM and HMM systems was compared using the same data under a forensically realistic, but challenging condition, which is sparsity of data. Even with short durations of two-, three-, four-, five- and seven-mora words, the HMM system constantly outperformed the GMM system in terms of average  $C_{llr}$  values. However, the benefits of the transitional information become evident when the HMM system is built with words longer than two- or three mora. With a seven-mora word, for example, the HMM system performed better than the GMM system by a  $C_{llr}$  value of 0.073.

This study also demonstrates that the outcomes (scores) of the GMM and HMM systems are not well-calibrated; thus calibration is an essential part of the FVC if they are to be used as models in the system.

## Acknowledgments

The authors thank the reviewers for their valuable comments.

## References

- Aitken, C. G. G. (1995). *Statistics and the Evaluation of Evidence for Forensic Scientists*. Chichester: John Wiley.
- Aitken, C. G. G., & Stoney, D. A. (1991). *The Use of Statistics in Forensic Science*. New York; London: Ellis Horwood.
- Aitken, C. G. G., & Taroni, F. (2004). *Statistics and the Evaluation of Evidence for Forensic Scientists*. Chichester: John Wiley & Sons.
- Balding, D. J., & Steele, C. D. (2015). *Weight-of-evidence for Forensic DNA Profiles*. Chichester: John Wiley & Sons.
- Brümmer, N., & du Preez, J. (2006). Application-independent evaluation of speaker detection. *Computer Speech and Language*, 20(2-3), 230-275.
- Burget, L., Plchot, O., Cumani, S., Glembek, O., Matějka, P., & Brümmer, N. (2011). Discriminatively trained probabilistic linear discriminant analysis for speaker verification. *Proceedings of the Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, 4832-4835.
- Enzinger, E., & Morrison, G. S. (2017). Empirical test of the performance of an acoustic-phonetic approach to forensic voice comparison under conditions similar to those of a real case. *Forensic Science International*, 277, 30-40.
- Enzinger, E., Morrison, G. S., & Ochoa, F. (2016). A demonstration of the application of the new paradigm for the evaluation of forensic evidence under conditions reflecting those of a real forensic-voice-comparison case. *Science & Justice*, 56(1), 42-57.
- Evett, I. W. (1998). Towards a uniform framework for reporting opinions in forensic science casework. *Science & Justice*, 38(3), 198-202.
- Gonzalez-Rodriguez, J., Rose, P., Ramos-Castro, D., Toledano, D. T., & Ortega-Garcia, J. (2007). Emulating DNA: Rigorous quantification of evidential weight in transparent and testable forensic speaker recognition. *Ieee Transactions on Audio Speech and Language Processing*, 15(7), 2104-2115.
- Kurematsu, A., Takeda, K., Sagisaka, Y., Katagiri, S., Kuwabara, H., & Shikano, K. (1990). At Japanese Speech Database as a Tool of Speech Recognition and Synthesis. *Speech Communication*, 9(4), 357-363.
- Makinae, H., Osanai, T., Kamada, T., & Tanimoto, M. (2007). *Construction and preliminary analysis of a large-scale bone-conducted speech database*. Proceedings of the Institute of Electronics, information and Communication Engineers, 97-102.
- Morrison, G. S. (2009). Forensic voice comparison and the paradigm shift. *Science & Justice*, 49(4), 298-308.
- Morrison, G. S. (2011). A comparison of procedures for the calculation of forensic likelihood ratios from acoustic-phonetic data: Multivariate kernel density (MVKD) versus Gaussian mixture model-universal background model (GMM-UBM). *Speech Communication*, 53(2), 242-256.
- Morrison, G. S. (2013). Tutorial on logistic-regression calibration and fusion: Converting a score to a likelihood ratio. *Australian Journal of Forensic Sciences*, 45(2), 173-197.
- Morrison, G. S. (2018). The impact in forensic voice comparison of lack of calibration and of mismatched conditions between the known-speaker recording and the relevant-population sample recordings. *Forensic Science International*, 283, E1-E7.
- Morrison, G. S., Enzinger, E., & Zhang, C. (2018). Forensic Speech Science. In I. Freckleton & H. Selby (Eds.), *Expert Evidence*. Sydney, Australia: Thomson Reuters.
- Morrison, G. S., Sahito, F. H., Jardine, G., Djokic, D., Clavet, S., Berghs, S., & Dorny, C. G. (2016). INTERPOL survey of the use of speaker identification by law enforcement agencies. *Forensic Science International*, 263, 92-100.
- Reynolds, D. A., Quatieri, T. F., & Dunn, R. B. (2000). Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, 10(1-3), 19-41.
- Robertson, B., & Vignaux, G. A. (1995). *Interpreting Evidence: Evaluating Forensic Science in the Courtroom*. Chichester: John Wiley.
- Rose, P. (2006). Technical forensic speaker recognition: Evaluation, types and testing of evidence. *Computer Speech and Language*, 20(2-3), 159-191.
- Rose, P. (2013). More is better: Likelihood ratio-based forensic voice comparison with vocalic segmental cepstra frontends. *International Journal of Speech, Language and the Law*, 20(1), 77-116.
- Rose, P. (2017). Likelihood ratio-based forensic voice comparison with higher level features: research and reality. *Computer Speech and Language*, 45, 475-502.

# Development of Natural Language Processing Tools for Cook Islands Māori

Rolando Coto-Solano<sup>1</sup>, Sally Akevai Nicholas<sup>2</sup>, and Samantha Wray<sup>3</sup>

<sup>1</sup>School of Linguistics and Applied Language Studies

Victoria University of Wellington

Te Whare Wānanga o te Īpoko o te Ika a Māui

[rolando.coto@vuw.ac.nz](mailto:rolando.coto@vuw.ac.nz)

<sup>2</sup>School of Language and Culture

Auckland University of Technology

[sally.nicholas@aut.ac.nz](mailto:sally.nicholas@aut.ac.nz)

<sup>3</sup>Neuroscience of Language Lab

New York University Abu Dhabi, United Arab Emirates

[samantha.wray@nyu.edu](mailto:samantha.wray@nyu.edu)

## Abstract

This paper presents three ongoing projects for NLP in Cook Islands Māori: Untrained Forced Alignment (approx. 9% error when detecting the center of words), automatic speech recognition (37% WER in the best trained models) and automatic part-of-speech tagging (92% accuracy for the best performing model). These new resources fill existing gaps in NLP for the language, including gold standard POS-tagged written corpora, transcribed speech corpora, and time-aligned corpora down to the phoneme level. These are part of efforts to accelerate the documentation of Cook Islands Māori and to increase its vitality amongst its users.

## 1 Introduction

Cook Islands Māori has been moderately well documented with two dictionaries ([Buse et al., 1996](#); [Savage, 1962](#)), a comprehensive description ([Nicholas, 2017](#)), and a corpus of audiovisual materials ([Nicholas, 2012](#)). However these materials are not yet sufficiently organized or annotated so as to be machine readable and thus maximally useful for both scholarship and revitalization projects. The human resources needed to achieve the desired level of annotation are not available, which has encouraged us to take advantage of NLP methods to accelerate documentation and research.

### 1.1 Minority Languages and NLP

Lack of resources makes it difficult to train data-driven NLP tools for smaller languages. This is compounded by the difficulty in generating input for Indigenous and endangered languages, where dwindling numbers of speakers, non-standardized writing systems and lack of resources to train specialist transcribers and analysts create a vicious cycle that makes it even more difficult to take advantage of NLP solutions. Amongst the hundreds of languages of the Americas, for example, very few have large spoken and written corpora (e.g. Zapotec from Mexico, Guaraní from Paraguay and Quechua from Bolivia and Perú), some have spoken and written corpora, and only a handful possess more sophisticated tools like spell-checkers and machine translation ([Mager et al., 2018](#)).

Overcoming these limitations is an important part of accelerating language documentation. Creating NLP resources also enhances the profile of endangered languages, creating a symbolic impact to generate positive associations towards the language and attract new learners, particularly young members of the community who might not otherwise see their language in a digital environment ([Aguilar Gil, 2014](#); [Kornai, 2013](#)).

As for Polynesian languages, te reo Māori, the Indigenous language spoken in Aotearoa New Zealand, is the one that has received the most attention from the NLP community. It has multiple corpora and Google provides machine-translations for it as part of *Google Translate*, and tools such as speech-to-text, text-to-speech and parsing are

under development (Bagnall et al., 2017). Other languages in the family have also received some attention. For example, Johnson et al., (2018) have worked on forced alignment for Tongan.

## 1.2 Cook Islands Māori

Southern Cook Islands Māori<sup>1</sup> (ISO 639-3 *rar*, or glottology *raro1241*) is an endangered East Polynesian language indigenous to the realm of New Zealand. It originates from the southern Cook Islands (see figure 1). Today however, most of its speakers reside in diaspora populations in Aotearoa New Zealand and Australia (Nicholas, 2018). The languages most closely related to Cook Islands Māori are the languages of Rakahanga/Manihiki (*rkh*, *raka1237*) and Penrhyn (*pnh*, *penr1237*) which originate from the northern Cook Islands. Te reo Māori (*mri*, *maor1246*) and Tahitian (*tah*, *tahi1242*), both East Polynesian languages, are also closely related. There is some degree of mutual intelligibility between these languages but they are generally considered to be separate languages by linguists and community members alike.

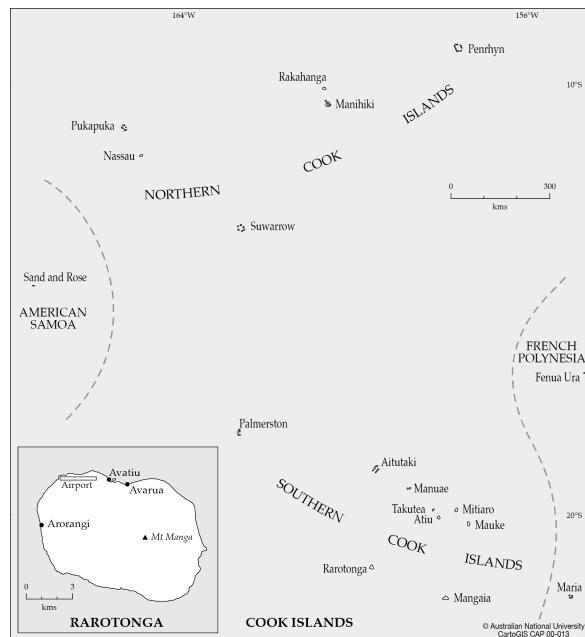


Figure 1: Cook Islands (CartoGIS Services et al., 2017)

Cook Islands Māori is severely endangered

<sup>1</sup> Southern Cook Islands Māori (henceforth Cook Islands Māori) has historically been called Rarotongan by non-Indigenous scholars. However, this name is disliked by the speech community and should not be used to refer to Southern Cook Islands Māori but rather to the specific variety originating from the island of Rarotonga (Nicholas, 2018:36).

(Nicholas, 2018, 46). Overall its vitality is between a 7 (shifting) and an 8a (moribund) on the Expanded Graded Intergenerational Disruption Scale (Lewis and Simons, 2010, 2). Among the diaspora population and that on the island of Rarotonga there has been a shift to English and there is very little intergenerational transmission of Cook Islands Māori. The only contexts where the vitality is strong is within the small populations of the remaining islands of the southern Cook Islands, where Cook Islands Māori still serves as the lingua franca (see Nicholas (2018) for a full discussion).

## 1.3 Grammatical Structure

The following is a selection of grammatical features of Cook Islands Māori as described in Nicholas (2017). Cook Islands Māori is of the isolating type with very few productive morphological processes. There is widespread polysemy, particularly within the grammatical particles. For example, in the following sentence, glossed using the Leipzig Glossing Rules (Bickel et al., 2008), there are four homophones of the particle *i*: the past tense marker, the cause preposition, the locative preposition and the temporal locative preposition.

- (1) *I mate a Mere i te mangō*  
 PST be-dead DET Mere CAUSE the shark  
*i roto i te moana i*  
 LOC inside LOC the ocean LOC.TIME  
*te Tapati.*  
 the Sunday

'Mere was killed by the shark in the ocean on Sunday.'

Furthermore, nearly every lexical item that can occur in a verb phrase can also occur in a noun phrase without any overt derivation, making tasks like POS tagging more difficult (see section 2.3). The unmarked constituent order is predicate initial. There are verbal and non-verbal predicate types. In sentences with verbal predicates the unmarked order is VSO. The phoneme paradigm is small, as is typical for Polynesian languages, with 9 consonants and 5 vowels which have a phonemic length distinction.

## 2 CIM NLP Projects Under Development

There is a need to accelerate the documentation of the endangered Cook Islands Māori language, so that it can be revitalized in Rarotonga and Aotearoa New Zealand and its usage domains can be expanded in the islands where it is still the lingua franca. We have begun working on this through three projects: (i) We have used untrained forced speech alignment to generate correspondences between transcriptions and their audio files, with the purpose of improving phonetic and phonological documentation. (ii) We are training speech-to-text models to automatize the transcription of both legacy material and recordings made during linguistic fieldwork. (iii) We are developing an interface-accompanied part-of-speech tagger as a first step towards full automatic parsing of the language. The following subsections provide details about the current state of each project.

### 2.1 Untrained Forced Alignment

Forced alignment is a technique that matches the sound wave with its transcription, down to the word and phoneme levels (Wightman and Talkin, 1997). This makes the work of generating alignment grids up to 30 times faster than manual processing (Labov et al., 2013). In theory, forced alignment needs a specific language model to function (e.g. an English language model aligning English text). However, untrained forced alignment, where for example Cook Islands Māori audio recordings and transcriptions are processed using an English language model, has been proven to be useful for the alignment of text and audio in Indigenous and under-resourced languages (Di-Canio et al., 2013).

In order to use this technique, a dictionary of Cook Islands Māori to English *Arpabet* was built, so that the Cook Islands Māori words could be introduced as new English words into the alignment tool. Some examples are shown in table 1. The phones of Cook Islands Māori words were matched with the closest English phone in the Arpabet system (e.g. the T in *kite* ‘to know’). Some Cook Islands Māori phones were not available in English; long Cook Islands Māori vowels were replaced by the equivalent accented vowel in English, and the glottal stop was replaced by the Arpabet phone T, as in *ngā‘i* ‘place’.

The English language acoustic model from FAVE-align (Rosenfelder et al., 2014) was used

CIM	Arpabet
kite	K IY1 T EH1
ngā‘i	NG AE1 T IY1

Table 1: Arpabet conversion of CIM data

to align previously transcribed recordings of Cook Islands Māori speech. Figure 2 shows the output of this process, a time-aligned transcription of the audio in the Praat (Boersma et al., 2002) TextGrid format.

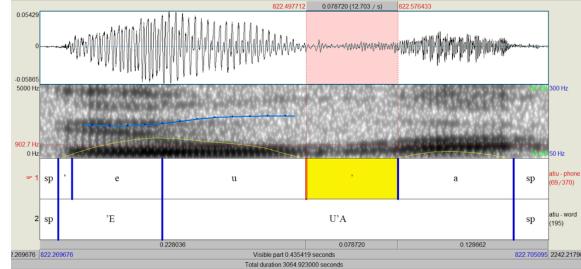


Figure 2: Praat TextGrid for CIM forced aligned text

After the automatic TextGrids were generated, 1045 phonemic tokens (628 vowels, 298 consonants, 119 glottal stops) and the words containing them were hand-corrected to verify the accuracy of the automatic system. Table 2 shows a summary of the results. The alignment showed error rates of 9% for the center of words and 20% for the center of vowels<sup>2</sup>. This error is higher than that observed for other instances of untrained forced alignment (2%, 7% and 3% than that observed for the Central American languages Bribri, Cabécar and Maleku respectively (Coto-Solano and Solorzano, 2016; Coto-Solano and Solórzano, 2017)), but it provides significant improvements in speed over hand alignment.

Type of interval	Error (relative to the duration of the interval)
Words	9% ± 12%
Vowels	20% ± 25%

Table 2: Errors for Untrained Forced Alignment

Utilization of forced alignment as a method for documentation and phonetic research has already

<sup>2</sup>We are currently documenting phonetic variation in vowels using data that was first force aligned and then manually corrected. Because of this, we don’t know at this point whether the 20% error is due to phonetic differences between English and CIM vowels, or if it’s due to the data itself.

been demonstrated for Austronesian languages. In [Nicholas and Coto-Solano \(2018\)](#), Cook Islands Māori phonemic tokens including vowels, consonants and glottal stops were forced-aligned and manually corrected to study the glottal stop phoneme. This work was able to show that in islands such as ‘Atiu, whose dialect is reported as having lost the glottal stop, the phoneme survives as shorter stop or as creaky voice. The corrected Praat TextGrids are publicly available at Paradisec ([Thieberger and Barwick, 2012](#)), in the collection of [Nicholas \(2012\)](#).

## 2.2 Automatic Speech Recognition

We have begun the training of an Automatic Speech Recognition (ASR) system using the Kaldi system ([Povey et al., 2011](#)), both independently and through the ELPIS pipeline developed by Co-EDL ([Foley et al., 2018](#)). While this work is still in progress, our preliminary results point to the need of custom models for each speaker. As can be seen in table 3, our recordings produced models with very different per-speaker word-error rates. (The “all speakers” model has cross-speaker test set). This, in addition to the paucity of data (approx. 80 minutes of speech for all speakers) makes the task extremely difficult.

Speaker	WER
Female, middle-aged, controlled environment	37%
Female, middle-aged, open environment	55%
Female, elderly, open environment	62%
Male, elderly, open environment	68%
All speakers	64%

Table 3: Errors for Untrained Forced Alignment (per-speaker)

The recording with the best performance was recorded in a very controlled environment (a silent room with a TASCAM DR-1 recorder). The worst recordings were those of elderly speakers who were speaking in their living rooms with open windows. The main issue here is that it is precisely these kinds of recordings (open environments with elderly practitioners of traditional knowledge or tellers of traditional stories) that are of most interest to linguists and practitioners of language re-

vitalization, and it is in those environments where we can see our worst performance. More work on this area is needed (e.g. crowdsourcing the recording of fixed phrases from numerous speakers for more reliable training).

## 2.3 Part-of-Speech Tagging

We have begun developing automatic part-of-speech (POS) tagging to aid in linguistic research of the syntax of Cook Islands Māori, and to build towards a full parser of the language. To begin our work we hand-tagged 418 sentences (2916 words) using the part of speech tags from [Nicholas \(2017\)](#). This is the only POS annotated corpus of Cook Islands Māori existing to date. The corpus is currently being prepared for public release.

The corpus is currently annotated using two levels of tagging: a more shallow/broad level with 23 tags, and a second, narrower level with 70 tags. For example, the shallow level contains tags like *v* for verbs, *n* for nouns and *t am* for tense-aspect-mood particles. The narrower level provides further detail for each tag. For example, it separates the verbs into *v:vt* for transitive verbs, *v:vi* for intransitives, *v:vstat* for stative verbs, and *v:vpass* for passives.

Classification experiments were carried out in the WEKA environment for machine learning ([Hall et al., 2009](#)). We tested algorithms that we believed would cope with the sparseness of the data given the size of the corpus. These algorithms were: *D. Tree* (*J48*): an open-source Java-based extension of the C4.5 decision tree algorithm ([Quinlan, 2014](#)) and *R. Forest*: merger of random decision trees (with 100 iterations). Included as a reference baseline is a *zeroR* classification algorithm predicting the majority POS class for all words. All algorithms were evaluated by splitting the entire corpus of 2916 words into a 90% set of sentences for training and 10% set for testing.

The model used position-dependent word context features for classification of each word’s POS. These included:

- the word ( $w$ )
- the previous word ( $w-1$ )
- the word before the previous word ( $w-2$ )
- the word two before the previous word ( $w-3$ )
- the following word ( $w+1$ )

Model	Accuracy
<i>Shallow/broad POS tags</i>	
D Tree (j48)	87.59%
R Forest (I = 100)	<b>92.41%</b>
zeroR (baseline)	21.03%
<i>Narrow POS tags</i>	
D Tree (j48)	80.00%
R Forest (I = 100)	82.41%
zeroR (baseline)	15.52%

Table 4: Accuracy of POS tagging models. Performance is reported in accuracy (per-token)

A comparison of the classification algorithms using the features above is shown in Table 4. As seen here, the current top performing classifier that we have identified is a *Random Forest* classifier. This algorithm performs best when the POS tags are less informative; that is, it performs best on the shallow/broad tags with an accuracy of 92.41%. Despite the fact that the narrow tags do not collapse across types and therefore are more difficult to classify, the best performer for the narrow tags is also the *Random Forest* classifier. This performance is comparable to other POS tagging tasks for under-resourced languages for which a new minimal dataset was manually tagged as the sole input for training, such as 71-78% for Kinyarwanda and Malagasy using a Hidden Markov Model (Garrette and Baldridge, 2013). It is also comparable to POS tagging for related languages with relatively larger corpora, such as Indonesian (94% accuracy, with 355,000 annotated tokens) (Fu et al., 2018).

To obtain an assessment of directions for future work aimed at improving the model, we also determined the most commonly confused tags by consulting a confusion matrix. The most common errors for the top performer (Shallow/broad tags as classified by the Random Forest) are seen in table 5. Recall from section 1.3 that grammatically, lexical items which occur in a noun phrase can also occur in a verb phrase with no overt derivational marking. This explains the fact that the majority of errors occurred as the result of confusion between v and n.

After training the model, we built a JavaServer Pages (JSP) interface to demo the model and obtain POS tagged versions of new, raw untagged sentences. This is illustrated in figure 3 below. The interface is in the process of being prepared

Error type assigned tag $\Rightarrow$ correct tag	% of errors
n (NOUN) $\Rightarrow$ v (VERB)	23%
tam (tense aspect mood) $\Rightarrow$ prep	9%
prep $\Rightarrow$ tam (tense aspect mood)	5%
v (VERB) $\Rightarrow$ n (NOUN)	5%

Table 5: Most common POS tagger errors for shallow/broad tags for top-performing tagger (Random Forest)

for public launch.

Figure 3: Interface for the POS tagger

The assignation of parts of speech is a very difficult task in Cook Islands Māori not only because of its data-driven nature, but because the orthography of Cook Islands Māori has not been fixed until recently. As detailed in Nicholas (2013), historically and even today there are numerous variations in spelling. The glottal stop is frequently omitted in spelling, as are the macrons for vocalic length. As a result, words like ‘e ‘nominal predicate marker’ can be written as ‘ē or e, increasing the homography of the language. Figure 4 shows the JSP interface attempting to tag two variants of the greeting ‘Aere mai ‘Welcome’. The first is spelled according to the current orthographic guidelines, with a glottal stop character at the beginning, and the first word gets tagged correctly as an intransitive verb. The second word, however, is spelled without the glottal stop, which leads the model to misidentify it as a noun. Future work includes experiments on how to best tackle this problem by utilizing naturally occurring variety written as spontaneous orthographies, looking at solutions already found for languages with non-standardized writing systems such as colloquial Arabic (Wray et al., 2015).



Figure 4: Error when tagging words without a glottal stop

## 2.4 Language Revitalization Applications

In addition to the scholarly advances these projects will support, they also have utility for the revitalization of Cook Islands Māori. The forced alignment work will support the teaching of the phonetics, phonology, and ‘pronunciation’, as well as improve community understanding of variation. The consumption of audio visual material with closed captions in the target language is known to be beneficial for language learning (Vanderplank, 2016), and automatic speech recognition will greatly increase the quantity of such material available to learners. The increased size of the searchable corpus of Cook Islands Māori will facilitate the production of corpus-based and multimedia pedagogical materials. Similarly, the POS tagged data will provide further benefits for the accurate design of pedagogical materials and the linguistic training of teachers who speak Cook Islands Māori. Additional tools that can be developed using this well annotated corpus include text-to-speech technology and chatbot applications.

## 3 Future work

There is much work to be done to move forward in these three projects. The next step for the POS tagging is to add resilience to cope with the non-standardized writing it will most commonly find. As for the speech recognition, crowdsourcing similar to that carried out by Bagnall et al (2017) might help increase the amount of controlled, pre-transcribed audio available for speech recognition training. Furthermore, we are investigating the incorporation of algorithms for treatment of audio data with low signal-to-noise ratio to improve audio quality which theoretically should lower the high word error rate for recordings conducted in an open, noisy environment.

## 4 Conclusions

This paper summarizes ongoing work in the application of NLP techniques to Cook Islands Māori, including untrained forced alignment for producing time-aligned transcriptions down to the phoneme, automatic speech recognition for the production of transcriptions from recordings of speech, and part-of-speech tagging for producing tagged text. We have given an overview of the state of these projects, and presented ideas for future work in this area. We believe that this interdisciplinary work can accelerate and enhance not only the documentation of the language, but can ultimately bring more of the Cook Islands community in contact with its language and help in its revitalization.

## Acknowledgments

We wish to thank Dr. Ben Foley and Dr. Miriam Meyerhoff of CoEDL for their support in numerous aspects of this project, as well as three anonymous reviewers for their helpful comments. We also wish to thank Jean Mason, Director of the Rarotonga Library, Teata Ateriano, principal of the School of Ma’uke, and Dr. Tyler Peterson from Arizona State University for their continued support in the documentation of Cook Islands Māori.

## References

- Yásnaya Elena Aguilar Gil. 2014. para qu publicar libros en lenguas indigenas si nadie los lee? e'px. <http://archivo.estepais.com/site/2014/paraquepublicar-libros-en-lenguas-indigenas-si-nadie-los-lee/>.
- Douglas Bagnall, Keoni Mahelona, and Peter-Lucas Jones. 2017. Kōrero Māori: a serious attempt at speech recognition for te reo Māori. Paper presented at the Conference 2017 NZ LingSoc, Auckland.
- Balthasar Bickel, Bernard Comrie, and Martin Haspelmath. 2008. The leipzig glossing rules. conventions for interlinear morpheme by morpheme glosses. *Revised version of February* .
- Paul Boersma et al. 2002. Praat, a system for doing phonetics by computer. *Glot international* 5.
- Jasper E Buse, Raututi Taringa, Bruce Biggs, and Rangi Moeka'a. 1996. *Cook Islands Māori dictionary with English-Cook Island Māori finderlist*. Pacific Linguistics, Research School of Pacific and Asian Studies, Australian National University, Canberra, ACT.

- CartoGIS Services, College of Asia and the Pacific, and The Australian National University. 2017. Cook islands. <http://asiapacific.anu.edu.au/mapsonline/base-maps/cook-islands> [Accessed 2017-10-06].
- Rolando Coto-Solano and Sofía Flores Solórzano. 2017. Comparison of two forced alignment systems for aligning bribri speech. *CLEI Electron. J.* 20(1):2–1.
- Rolando Coto-Solano and Sofia Flores Solorzano. 2016. Alineación forzada sin entrenamiento para la anotación automática de corpus orales de las lenguas indígenas de costa rica. *Kāñina* 40(4):175–199.
- Christian DiCanio, Hosung Nam, Douglas H Whalen, H Timothy Bunnell, Jonathan D Amith, and Rey Castillo García. 2013. Using automatic alignment to analyze endangered language data: Testing the viability of untrained alignment. *The Journal of the Acoustical Society of America* 134(3):2235–2246.
- Ben Foley, Josh Arnold, Rolando Coto-Solano, Gauthier Durantin, T. Mark Ellison, Daan van Esch, Scott Heath, Frantisek Kratochvil, Zara Maxwell-Smith, David Nash, Ola Olsson, Mark Richards, Nay San, Hywel Stoakes, Nick Thieberger, and Janet Wiles. 2018. Building speech recognition systems for language documentation: The coedl endangered language pipeline and inference system (elpis). In *6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages*.
- Sihui Fu, Nankai Lin, Gangqin Zhu, and Shengyi Jiang. 2018. Towards indonesian part-of-speech tagging: Corpus and models. In *LREC*.
- Dan Garrette and Jason Baldridge. 2013. Learning a part-of-speech tagger from two hours of annotation. In *Proceedings of NAACL 2013*. Atlanta, USA, pages 138–147.
- Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. 2009. The weka data mining software: an update. *ACM SIGKDD explorations newsletter* 11(1):10–18.
- Lisa M Johnson, Marianna Di Paolo, and Adrian Bell. 2018. Forced alignment for understudied language varieties: Testing prosodylab-aligner with tongan data. *Language Documentation & Conservation*.
- András Kornai. 2013. Digital language death. *Plos one* 8(10):e77056.
- William Labov, Ingrid Rosenfelder, and Josef Fruehwald. 2013. One hundred years of sound change in philadelphia: Linear incrementation, reversal, and reanalysis. *Language* 89(1):30–65.
- M. Paul Lewis and Gary F. Simons. 2010. Making EGIDS assessment for the ethnologue. [www.icc.org.kh/download/Making\\_EGIDS-Assessments\\_English.pdf](http://www.icc.org.kh/download/Making_EGIDS-Assessments_English.pdf) [Accessed 2017-07-20].
- Manuel Mager, Ximena Gutierrez-Vasques, Gerardo Sierra, and Ivan Meza. 2018. Challenges of language technologies for the indigenous languages of the americas. *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018)*.
- Sally Nicholas. 2017. *Ko te Karāma o te Reo Māori o te Pae Tonga o Te Kuki Airani: A Grammar of Southern Cook Islands Māori*. Ph.D. thesis, University of Auckland.
- Sally Akevai Nicholas. 2012. *Te Vairanga Tuatua o te Te Reo Māori o te Pae Tonga: Cook Islands Māori (Southern dialects) (sn1)*. Digital collection managed by PARADISEC. [Open Access] DOI: 10.4225/72/56E9793466307 <http://catalog.paradisec.org.au/collections/SN1>.
- Sally Akevai Nicholas. 2018. Language contexts: Te Reo Māori o te Pae Tonga o te Kuki Airani also known as Southern Cook Islands Māori. *Language Documentation and Description* 15:36–64.
- Sally Akevai Nicholas and Rolando Coto-Solano. 2018. Using untrained forced alignment to study variation of glottalization in cook islands mori. In *NWAV-AP5*. University of Brisbane.
- Sally Akevai Te Namu Nicholas. 2013. Orthographic reform in cook islands māori: Human considerations and language revitalisation implications. *3rd International Conference on Language Documentation and Conservation (ICLDC)*.
- Daniel Povey, Arnab Ghoshal, Gilles Boulian, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. 2011. The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, EPFL-CONF-192584.
- J Ross Quinlan. 2014. *C4. 5: programs for machine learning*. Elsevier.
- Ingrid Rosenfelder, Josef Fruehwald, Keelan Evanini, Scott Seyfarth, Kyle Gorman, Hilary Prichard, and Jiahong Yuan. 2014. FAVE (Forced Alignment and Vowel Extraction) Program Suite.
- Stephen Savage. 1962. *A dictionary of the Maori language of Rarotonga / manuscript by Stephen Savage*. New Zealand Department of Island Territories, Wellington.
- Nicholas Thieberger and Linda Barwick. 2012. Keeping records of language diversity in melanesia, the pacific and regional archive for digital sources in endangered cultures (paradisec). *Melanesian languages on the edge of Asia: Challenges for the 21st Century* pages 239–53.

Robert Vanderplank. 2016. Effects of and effects with captions: How exactly does watching a tv programme with same-language subtitles make a difference to language learners?. *Language Teaching* 49(2):235.

Colin W Wightman and David T Talkin. 1997. The aligner: Text-to-speech alignment using markov models. In *Progress in speech synthesis*, Springer, pages 313–323.

Samantha Wray, Hamdy Mubarak, and Ahmed Ali. 2015. Best Practices for Crowdsourcing Dialectal Arabic Speech Transcription. In *Proceedings of Workshop on Arabic Natural Language Processing*.

# Unsupervised Mining of Analogical Frames by Constraint Satisfaction

Lance De Vine

Shlomo Geva

Peter Bruza

Queensland University of Technology

Brisbane, Australia

{l.devine, s.geva, p.bruza}@qut.edu.au

## Abstract

It has been demonstrated that vector-based representations of words trained on large text corpora encode linguistic regularities that may be exploited via the use of vector space arithmetic. This capability has been extensively explored and is generally measured via tasks which involve the automated completion of linguistic proportional analogies. The question remains, however, as to what extent it is possible to induce relations from word embeddings in a principled and systematic way, without the provision of exemplars or seed terms. In this paper we propose an extensible and efficient framework for inducing relations via the use of constraint satisfaction. The method is efficient, unsupervised and can be customized in various ways. We provide both quantitative and qualitative analysis of the results.

## 1 Introduction

The use and study of analogical inference and structure has a long history in linguistics, logic, cognitive psychology, scientific reasoning and education (Bartha, 2016), amongst others. The use of analogy has played an especially important role in the study of language, language change, and language acquisition and learning (Kiparsky, 1992). It has been a part of the study of phonology, morphology, orthography and syntactic grammar (Skousen, 1989), as well as the development of applications such as machine translation and paraphasing (Lepage and Denoual, 2005).

Recent progress with the construction of vector space representations of words based on their distributional profiles has revealed that analogical structure can be discovered and operationalised via the use of vector space algebra (Mikolov et al., 2013b). There remain many questions regarding the extent to which word vectors encode analogical structure and also the extent to which this

structure can be uncovered. For example, we are not aware of any proposal or system that is focussed on the unsupervised and systematic discovery of analogies from word vectors that does not make use of exemplar relations, existing linguistic resources or seed terms. The automatic identification of linguistic analogies, however, offers many potential benefits for a diverse range of research and applications, including language learning and computational creativity.

Computational models of analogy have been studied since at least the 1960’s (Hall, 1989; French, 2002) and have addressed tasks relating to both proportional and structural analogy. Many computational systems are built as part of investigations into how humans might perform analogical inference (Gentner and Forbus, 2011). Most make use of Structure Mapping Theory (SMT) (Gentner, 1983) or a variation thereof which maps one relational system to another, generally using a symbolic representation. Other systems use vector space representations constructed from corpora of natural language text (Turney, 2013). The analogies that are computed using word embeddings have primarily been proportional analogies and are closely associated with the prediction of relations between words. For example, a valid semantic proportional analogy is “cat is to feline as dog is to canine” which can be written as “cat : feline :: dog : canine.”

In linguistics proportional analogies have been extensively studied in the context of both inflectional and derivational morphology (Blevins, 2016). Proportional analogies are used as part of an inference process to fill the cells/slots in a word *paradigm*. A paradigm is an array of morphological variations of a lexeme. For example, {cat, cats} is a simple singular-noun, plural-noun paradigm in English. Word paradigms exhibit inter-dependencies that facilitate the inference of

new forms and for this reason have been studied within the context of language change. The informativeness of a form correlates with the degree to which knowledge of the form reduces uncertainty about other forms within the same paradigm (Blevins et al., 2017).

In this paper we propose a construction which we call an *analogical frame*. It is intended to elicit associations with the terms *semantic frame* and *proportional analogy*. It is an extension of a linguistic analogy in which the elements satisfy certain constraints that allow them to be induced in an unsupervised manner from natural language text. We expect that analogical frames will be useful for a variety of purposes relating to the automated induction of syntactic and semantic relations and categories .

The primary contributions of this paper are two-fold:

1. We introduce a generalization of proportional analogies with word embeddings which we call *analogical frames*.
2. We introduce an efficient constraint satisfaction based approach to inducing analogical frames from natural language embeddings in an unsupervised fashion.

In section 2 we present background and related research. In section 3 we present and explain the proposal of Analogical Frames. In section 4 we present methods implemented for ensuring search efficiency of Analogical Frames. In section 5 we present some analysis of empirical results. In section 6 we present discussion of the proposal and in section 7 we conclude.

## 2 Background and Related Work

### 2.1 Proportional Analogies

A proportional analogy is a 4-tuple which we write as  $x_1 : x_2 :: x_3 : x_4$  and read as “ $x_1$  is to  $x_2$  as  $x_3$  is to  $x_4$ ”, with the elements of the analogy belonging to some domain  $X$  (we use this notation as it is helpful later). From here-on we will use the term “analogy” to refer to proportional analogies unless indicated otherwise. Analogies can be defined over different types of domains, for example, strings, geometric figures, numbers, vector spaces, images etc. (Stroppa and Yvon, 2005) propose a definition of proportional analogy over any domain which is equipped with an internal

composition law  $\oplus$  making it a semi-group  $(X, \oplus)$ . This definition also applies to any richer algebraic structure such as groups or vector spaces. In  $\mathbb{R}^n$ , given  $x_1, x_2$  and  $x_3$  there is always only one point that can be assigned to  $x_4$  such that proportionality holds. (Miclet et al., 2008) define a relaxed form of analogy which reads as “ $x_1$  is to  $x_2$  almost as  $x_3$  is to  $x_4$ ”. To accompany this they introduce a measure of *analogical dissimilarity* (AD) which is a positive real value and takes the value 0 when the analogy holds perfectly. A set of four points  $\mathbb{R}^n$  can therefore be scored for analogical dissimilarity and ranked.

### 2.2 Word Vectors and Proportional Analogies

The background just mentioned provides a useful context within which to place the work on linguistic regularities in word vectors (Mikolov et al., 2013b; Levy et al., 2014). (Mikolov et al., 2013b) showed that analogies can be completed using vector addition of word embeddings. This means that given  $x_1, x_2$  and  $x_3$  it is possible to infer the value of  $x_4$ . This is accomplished with the vector offset formula, or 3CosAdd (Levy et al., 2014).

$$\arg \max_{x_4} s(x_4, x_2 + x_3 - x_1) \quad \text{3CosAdd}$$

The  $s$  in 3CosAdd is a similarity measure. In practice unit vectors are generally used with cosine similarity. (Levy et al., 2014) introduced an expression 3CosMul which tends to give a small improvement when evaluated on analogy completion tasks.

$$\arg \max_{x_4} \frac{s(x_4, x_3) \cdot s(x_4, x_2)}{s(x_4, x_1) + \epsilon} \quad \text{3CosMul}$$

3CosAdd and 3CosMul are effectively scoring functions that are used to judge the correctness of a value for  $x_4$  given values for  $x_1, x_2$  and  $x_3$ .

### 2.3 Finding Analogies

Given that it is possible to complete analogies with word vectors it is natural to ask whether analogies can be identified without being given  $x_1, x_2$  and  $x_3$ . (Stroppa and Yvon, 2005) considers an analogy to be valid when analogical proportions hold between all terms in the analogy. They describe a finite-state solver which searches for formal analogies in the domain of strings and trees. As noted by several authors (Lavallée and Langlais, 2009;

(Langlais, 2016; Beltran et al., 2015) a brute force search to discovering proportional analogies is computationally difficult, with the complexity of a naive approach being at least  $O(n^3)$  and perhaps  $O(n^4)$ . A computational procedure must at least traverse the space of all 3-tuples if assuming that the 4th term of an analogy can be efficiently inferred.

For example, if we use a brute force approach to discovering linguistic analogies using a vocabulary of 100 000 words, we would need to examine all combinations of 4-tuples, or 100000<sup>4</sup>, or 10<sup>20</sup> combinations. Various strategies may be

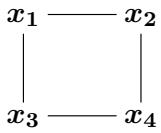


Figure 1: A Proportional Analogy template, with 4 variables to be assigned appropriate values

considered for making this problem tractable. The first observation is that symmetries of an analogy should not be recomputed (Lepage, 2014). This can reduce the compute time by 8 for a single analogy as there are 8 symmetric configurations.

Another proposed strategy is the construction of a feature tree for the rapid computation and analysis of tuple differences over vectors with binary attributes (Lepage, 2014). This method was used to discover analogies between images of Chinese characters. This has complexity  $O(n^2)$ . It computes the  $\frac{n(n+1)}{2}$  vectors between pairs of tuples and collects them together into clusters containing the same difference vector. A variation of this method was reported in (Fam and Lepage, 2016) and (Fam and Lepage, 2017) for automatically discovering *analogical grids* of word paradigms using edit distances between word strings. It is not immediately obvious, however, how to extend this to the case of word embeddings where differences between word representations are real valued vectors.

A related method is used in (Beltran et al., 2015) for identifying analogies in relational databases. It is less constrained as it uses analogical dissimilarity as a metric when determining valid analogies.

(Langlais, 2016) extend methods from (Lepage,

2014) to scale to larger datasets for the purpose of machine translation, but also limit themselves to the formal or graphemic level instead of more general semantic relations between words.

## 2.4 Other Related Work

The present work is related to a number of research themes in language learning, relational learning and natural language processing. We provide a small sample of these.

(Holyoak and Thagard, 1989) introduce the use of constraint satisfaction as a key requirement for models of analogical mapping. Their computer program ACME (Analogical Constraint Mapping Engine) uses a connectionist network to balance structural, semantic and pragmatic constraints for mapping relations. (Hummel and Holyoak, 1997) propose a computational model of analogical inference and schema induction using distributed patterns for representing objects and predicates. (Doumas et al., 2008) propose a computational model which provides an account of how structured relation representations can be learned from unstructured data. More specifically to language acquisition, (Bod, 2009) uses analogies over trees to derive and analyse new sentences by combining fragments of previously seen sentences. The proposed framework is able to replicate a range of phenomena in language acquisition.

From a more cognitive perspective (Kurtz et al., 2001) investigates how mutual alignment of two situations can create better understanding of both. Related to this (Gentner, 2010), and many others, argue that analogical ability is the key factor in human cognitive development.

(Turney, 2006, 2013) makes extensive investigations of the use of corpus based methods for determining relational similarities and predicting analogies.

(Miclet and Nicolas, 2015) propose the concept of an analogical complex which is a blend of analogical proportions and formal concept analysis.

More specifically in relation to word embeddings, (Zhang et al., 2016) presents an unsupervised approach for explaining the meaning of terms via word vector comparison.

In the next section we describe an approach which addresses the task of inducing analogies in an unsupervised fashion from word vectors and builds on existing work relating to word embeddings and linguistic regularities.

### 3 Analogical Frames

The primary task that we address in this paper is the discovery of linguistic proportional analogies in an unsupervised fashion given only the distributional profile of words. The approach that we take is to consider the problem as a constraint satisfaction problem (CSP) (Rossi et al., 2006). We increase the strength of constraints until we can accurately decide when a given set of words and their embeddings forms a valid proportional analogy. At this point we introduce some terminology:

Constraint satisfaction problems are generally defined as a triple  $P = \langle X, D, C \rangle$  where  
 $X = \{x_1, \dots, x_n\}$  is a set of variables.  
 $D = \{d_1, \dots, d_n\}$  is the set of domains associated with the variables.  
 $C = \{c_1, \dots, c_m\}$  is the set of constraints to be satisfied.

A solution to a constraint satisfaction problem must assign a single value to each variable  $x_1, \dots, x_n$  in the problem. There may be multiple solutions. In our problem formulation there is only one domain which is the set of word types which comprise the vocabulary. Each variable must be assigned a word identifier and each word is associated with one or more vector space representations. Constraints on the words and associated vector space representation limit the values that the variables can take. From here-on we will use the bolded symbol  $\boldsymbol{x}_i$  to indicate the vector space value of the word assigned to the variable  $x_i$ .

In our proposal we use the following five constraints.

*C1. AllDiff constraint.* The *Alldiff* constraint constrains all terms of the analogy to be distinct (The *Alldiff* constraint is a common constraint in CSPs), such that  $x_i \neq x_j$  for all  $1 < i < j < n$ .

*C2. Asymmetry constraint.* The *asymmetry constraint* (Meseguer and Torras, 2001) is used to eliminate unnecessary searches in the search tree. It is defined as a partial ordering on the values of a subset of the variables. In the case of a 2x3 analogical frame (figure 2), for example, we define the ordering as:

$$x_1 \prec x_2 \prec x_3, \text{ and } x_1 \prec x_4.$$

where the ordering is defined on the integer identifiers of the words in the vocabulary.

*C3. Neighbourhood Constraint.* The *neighbourhood constraint* is used to constrain the value

of variables to the words which are within the nearest neighbourhood of the words to which they are connected. We define this as:

$$x_i \in \text{Neigh}_t(x_j) \text{ and } x_j \in \text{Neigh}_t(x_i)$$

where  $\text{Neigh}_t(x_i)$  is the nearest neighbourhood of  $t$  words of the word assigned to  $x_i$ , as measured in the vector space representation of the words.

*C4. Parallel Constraint.* The *Parallel Constraint* forces opposing difference vectors to have a minimal degree of parallelism.

$$\widehat{\boldsymbol{x}_2 - \boldsymbol{x}_1} \cdot \widehat{\boldsymbol{x}_5 - \boldsymbol{x}_4} < pThreshold$$

where  $pThreshold$  is a parameter. For the parallel constraint we ensure that the difference vector  $\boldsymbol{x}_2 - \boldsymbol{x}_1$  has a minimal cosine similarity to the difference vector  $\boldsymbol{x}_5 - \boldsymbol{x}_4$ . This constraint overlaps to some extent with the proportionality constraint (below), however it serves a different purpose, which is to eliminate low probability candidate analogies.

*C5. Proportionality Constraint.* The *Proportionality Constraint* constrains the vector space representation of words to form approximate geometric proportional analogies. It is a quaternary constraint. For any given 4-tuple, we use the concept of “inter-predictability” (Blevins et al., 2017) to decide whether the 4-tuple is acceptable. We enforce inter-predictability by requiring that each term in a 4-tuple is predicted by the other three terms. This implies four analogy completion tasks which must be satisfied for the variable assignment to be accepted.

$$\boldsymbol{x}_1 : \boldsymbol{x}_2 :: \boldsymbol{x}_3 : \boldsymbol{x} \Rightarrow \boldsymbol{x} = \boldsymbol{x}_4$$

$$\boldsymbol{x}_2 : \boldsymbol{x}_1 :: \boldsymbol{x}_4 : \boldsymbol{x} \Rightarrow \boldsymbol{x} = \boldsymbol{x}_3$$

$$\boldsymbol{x}_3 : \boldsymbol{x}_4 :: \boldsymbol{x}_1 : \boldsymbol{x} \Rightarrow \boldsymbol{x} = \boldsymbol{x}_2$$

$$\boldsymbol{x}_4 : \boldsymbol{x}_3 :: \boldsymbol{x}_2 : \boldsymbol{x} \Rightarrow \boldsymbol{x} = \boldsymbol{x}_1$$

(Stroppa and Yvon, 2005) use a similar approach with exact formal analogies. With our approach, however, we complete analogies using word vectors and analogy completion formulas (eg. using 3COSADD or 3CosMUL, or a derivative).

The proportionality constraint is a relatively expensive constraint to enforce as it requires many vector-vector operations and comparison against all word vectors in the vocabulary. This constraint may be checked approximately which we discuss in the next section.

### 3.1 The Insufficiency of 2x2 Analogies

When we experiment with discovering 2x2 analogies using the constraints just described we find that we can't easily set the parameters of the constraints such that only valid analogies are produced, without severely limiting the types of analogies which we accept. For example, the following analogy is produced "oldest : old :: earliest : earlier". We find that these types of mistakes are common. We therefore take the step of expanding the number of variables so that the state space is now larger (figure 2).

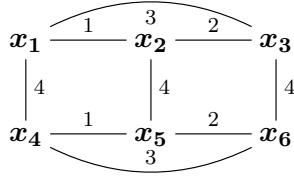


Figure 2: A 2x3 Analogical Frame

The idea is to increase the inductive support for each discovered analogy by requiring that analogies be part of a larger system of analogies. We refer to the larger system of analogies as an *Analogical Frame*. It is important to note that in figure 2,  $x_1$  and  $x_3$  are connected. The numbers associated with the edges indicate aligned vector differences. It is intended that the analogical proportions hold according to the connectivity shown. For example, proportionality should hold such that  $x_1 : x_2 :: x_4 : x_5$  and  $x_1 : x_3 :: x_4 : x_6$  and  $x_2 : x_3 :: x_5 : x_6$ . It is also important to note that while we have added only two new variables, we have increased the number of constraints by almost 3 times. It is not exactly 3 because there is some redundancy in the proportions.

### 3.2 New Formulas

We now define a modified formula for completing analogies that are part of a larger systems of analogies such as the analogical frame in figure 2. It can be observed that 3CosADD and 3CosMUL effectively assigns a score to vocabulary items and then selects the item with the largest score. We do the same but with a modified scoring function which is a hybrid of 3CosADD and 3COSMUL. 3COSADD or 3COSMUL could both be used as part of our approach, but the formula which we propose, better captures the intuition of larger systems of analogies where the importance is placed on 1) symmetry, and 2) average offset vectors.

When we are not given any prior knowledge, or exemplar, there is no privileged direction within an analogical frame. For example, in figure 2, the pair  $(x_2, x_5)$  has the same importance as  $(x_4, x_5)$ .

We first construct difference vectors and then average those that are aligned.

$$\begin{aligned} \text{dif}_{2,1} &= \widehat{x_2 - x_1} & \text{dif}_{4,1} &= \widehat{x_4 - x_1} \\ \text{dif}_{5,4} &= \widehat{x_5 - x_4} & \text{dif}_{5,2} &= \widehat{x_5 - x_2} \\ \text{dif}_{6,3} &= \widehat{x_6 - x_3} & \text{dif}_{6,3} &= \widehat{x_6 - x_3} \\ \text{difsum}_1 &= \text{dif}_{2,1} + \text{dif}_{5,4} \\ \text{difsum}_2 &= \text{dif}_{4,1} + \text{dif}_{5,2} + \text{dif}_{6,3} \\ \text{dif}_1 &= \frac{\text{difsum}_1}{|\text{difsum}_1|} \\ \text{dif}_2 &= \frac{\text{difsum}_2}{|\text{difsum}_2|} \end{aligned}$$

The vector  $\text{dif}_1$  is the normalized average offset vector indicated with a 1 in figure 2. The vector  $\text{dif}_2$  is the normalized average offset vector indicated with a 4 in figure 2.

Using these normalized average difference vectors we define the scoring function for selecting  $x_5$  given  $x_1, x_2$  and  $x_4$  as:

$$\begin{aligned} \arg \max_{x_5} s(x_5, x_4) \cdot s(x_5, \text{dif}_1) &\quad (1) \\ &+ s(x_5, x_2) \cdot s(x_5, \text{dif}_2) \end{aligned}$$

The formulation makes use of all information available in the analogical frame. Previous work has provided much evidence for the linear compositionality of word vectors as embodied by 3CosADD (Vylomova et al., 2015; Hakami et al., 2017). It has also been known since (Mikolov et al., 2013a) that averaging the difference vectors of pairs exhibiting the same relation results in a difference vector with better predictive power for that relation (Drozd et al., 2016).

Extrapolating from figure 2 larger analogical frames can be constructed, such as 2 x 4, 3 x 3, or 2 x 2 x 3. Each appropriately connected 4-tuple contained within the frame should satisfy the proportionality constraint.

## 4 Frame Discovery and Search Efficiency

The primary challenge in this proposal is to efficiently search the space of variable assignments. We use a depth first approach to cover the search space as well several other strategies to make this search efficient.

## 4.1 Word Embedding Neighbourhoods

The most important strategy is the concentration of compute resources on the nearest neighbourhoods of word embeddings. Most analogies involve terms that are within the nearest neighbourhood of each other when ranked according to similarity (Linzen, 2016). We therefore compute the nearest neighbour graph of every term in the vocabulary as a pre-processing step and store the result as an adjacency list for each vocabulary item. We do this efficiently by using binary representations of the word embeddings (Jurgovsky et al., 2016) and re-ranking using the full precision word vectors. The nearest neighbour list of each vocabulary entry is used in two different ways, 1) for exploring the search space of variable assignments, and 2) efficiently eliminating variable assignments (next section) that do not satisfy the proportional analogy constraints.

When traversing the search tree we use the adjacency list of an already assigned variable to assign a value to a nearby variable in the frame. The breadth of the search tree at each node is therefore parameterized by a global parameter  $t$  assigned by the user. The parameter  $t$  is one of the primary parameters of the algorithm and will determine the length of the search and the size of the set of discovered frames.

## 4.2 Elimination of Candidates by Sampling

A general principle used in most CSP solving is the quick elimination of improbable solutions. When we assign values to variables in a frame we need to make sure that the proportional analogy constraint is satisfied. For four given variable assignments  $w_1, w_2, w_3$  and  $w_4$  this involves making sure that  $w_4$  is the best candidate for completing the proportional analogy involving  $w_1, w_2$  and  $w_3$ . This is equivalent to knowing if there are any better vocabulary entries than  $w_4$  for completing the analogy. Instead of checking all vocabulary entries we can limit the list of entries to the  $m$  nearest neighbours of  $w_2, w_3$  and  $w_4$ <sup>1</sup> to check if any score higher than  $w_4$ . If any of them score higher the proportional analogy constraint is violated and the variable assignment is discarded. The advantage of this is that we only need to check  $3 \times m$  entries to approximately test the correctness of  $w_4$ .

<sup>1</sup>assuming  $w_1$  is farthest away from  $w_4$  and that the nearest neighbours of  $w_1$  are not likely to help check the correctness of  $w_4$

We find that if we set  $m$  to 10 then most incorrect analogies are eliminated. An exhaustive check for correctness can be made as a post processing step.

## 4.3 Other Methods

Another important method for increasing efficiency is testing the degree to which opposing difference vectors in a candidate proportional analogy are parallel to each other. This is encapsulated in constraint C4. While using parallelism as a parameter for solving word analogy problems has not been successful (Levy et al., 2014), we have found that the degree of parallelism is a good indicator of the confidence that we can have in the analogy once other constraints have been satisfied. Other methods employed to improve efficiency include the use of Bloom filters to test neighbourhood membership and the indexing of discovered proportions so as not to repeat searching.

## 4.4 Extending Frames

Frames can be extended by extending the initial *base frame* in one or more directions and searching for additional variable assignments that satisfy all constraints. For example, a 2x3 frame can be extended to a 2x4 frame by assigning values to two new variables  $x_7$  and  $x_8$  (figure 3).

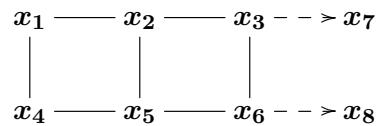


Figure 3: Extending a 2x3 frame

As frames are extended the average offset vectors (equation 1) are recomputed so that the offset vectors become better predictors for computing proportional analogies.

## 5 Experimental Setup and Results

The two criteria we use to measure our proposal include a) accuracy of analogy discovery, b) compute scalability. Other criteria are also possible such as relation diversity and interestingness of relations.

The primary algorithm parameters are 1) The number of terms in the vocabulary to search, 2) the size of the nearest neighbourhood of each term to search, 3) the degree of parallelism required for opposing vector differences and 4) whether to extend base frames.

## 5.1 Retrieving Analogical Completions from Frames

When frames are discovered they are stored in a *Frame Store*, a data structure used to efficiently store and retrieve frames. Frames are indexed using posting lists similar to a document index. To complete an incomplete analogy, all frames which contain all terms in the incomplete analogy are retrieved. For each retrieved frame the candidate term for completing the analogy is determined by cross-referencing the indices of the terms within the frame (figure 4). If diverse candidate terms are selected from multiple frames, voting is used to select the final analogy completion, or random selection in the case of tied counts.

$$a_{1,1} : a_{1,3} :: a_{3,1} : ?$$

<b>a<sub>1,1</sub></b>	<i>a<sub>1,2</sub></i>	<b>a<sub>1,3</sub></b>
<i>a<sub>2,1</sub></i>	<i>a<sub>2,2</sub></i>	<i>a<sub>2,3</sub></i>
<b>a<sub>3,1</sub></b>	<i>a<sub>3,2</sub></i>	<u><i>a<sub>3,3</sub></i></u>

Figure 4: Determining an analogy completion from a larger frame

We conducted experiments with embeddings constructed by ourselves as well as with publicly accessible embeddings from the fastText web site<sup>2</sup> trained on 600B tokens of the Common Crawl (Mikolov et al., 2018).

We evaluated the accuracy of the frames by attempting to complete the analogies from the well known Google analogy test set.<sup>3</sup> The greatest challenge in this type of evaluation is adequately covering the evaluation items. At least three of the terms in an analogy completion item need to be simultaneously present in a single frame for the item to be attempted. We report the results of a typical execution of the system using a nearest neighbourhood size of 25, a maximum vocabulary size of 50 000, and a minimal cosine similarity between opposing vector differences of 0.3. For this set of parameters, 8589 evaluation items were answered by retrieval from the frame store, covering approximately 44% of the evaluation items (table 1). Approximately 30% of these were from the semantic category, and 70% from the syntactic. We compared the accuracy of completing analogies using the frame store, to the accuracy of both 3CosAdd

Table 1: Analogy Completion on Google Subset

	3CosAdd	3CosMul	Frames
Sem. (2681)	2666	2660	<b>2673</b>
Syn. (5908)	5565	5602	<b>5655</b>
Tot. (8589)	8231	8262	<b>8328</b>

and 3CosMul using the same embeddings as used to build the frames.

Results show that the frames are slightly more accurate than 3CosAdd and 3CosMul, achieving 96.9% on the 8589 evaluation items. It needs to be stressed, however, that the objective is not to outperform vector arithmetic based methods, but rather to verify that the frames have a high degree of accuracy.

To better determine the accuracy of the discovered frames we also randomly sampled 1000 of the 21571 frames generated for the results shown in table 2, and manually checked them. The raw outputs are included in the online repository<sup>4</sup>. These frames cover many relations not included in the Google analogy test set. We found 9 frames with errors giving an accuracy of 99.1%.

It should be noted that the accuracy of frames is influenced by the quality of the embeddings. However, even with embeddings trained on small corpora it is possible to discover analogies provided that sufficient word embedding training epochs have been completed.

The online repository contains further empirical evaluations and explanations regarding parameter choices, including raw outputs and errors made by the system.

## 5.2 Scaling: Effect of Neighbourhood Size and pThreshold

Tables 2 and 3 show the number of frames discovered on typical executions of the software. The reported numbers are intended to give an indication of the relationship between neighbourhood size, number of frames produced and execution time. The reported times are for 8 software threads.

## 5.3 Qualitative Analysis

From inspection of the frames we see that a large part of the relations discovered are grammatical or morpho-syntactic, or are related to high frequency

<sup>2</sup><https://fasttext.cc/docs/en/english-vectors.html>

<sup>3</sup><http://download.tensorflow.org/data/questions-words.txt>

<sup>4</sup><https://github.com/ldevine/AFM>

Table 2: Par Thres = 0.3 (Less Constrained)

Near. Sz.	15	20	25	30
Frames	13282	16785	19603	21571
Time (sec)	20.1	29.3	38.3	45.9

Table 3: Par Thres = 0.5 (More Constrained)

Near. Sz.	15	20	25	30
Frames	6344	7995	9286	10188
Time (sec)	10.4	15.7	21.1	26.0

entities. However we also observe a large number of other types of relations such as synonyms, antonyms, domain alignments and various syntactic mappings. We provide but a small sample below.

eyes	eye	blindness
ears	ear	deafness
Father	Son	Himself
Mother	Daughter	Herself
geschichte	gesellschaft	musik
histoire	societe	musique
aircraft	flew	skies
ships	sailed	seas
		air
always	constantly	constant
often	frequently	frequent
sometimes	occasionally	occasional
brighter	stronger	
bright	+	strong
	+	
fainter	weaker	
faint	+	weak

(2 x 2 x 2 Frame)

We have also observed reliable mappings with embeddings trained on non-English corpora.

The geometry of analogical frames can also be explored via visualizations of projections onto 2D sub-spaces derived from the offset vectors.<sup>5</sup>

## 6 Discussion

We believe that the constraint satisfaction approach introduced in this paper is advantageous because it is a systematic but flexible and can make use of methods from the constraint satisfaction domain. We have only mentioned a few of the primary CSP concepts in this paper. Other constraints can be included in the formulation such

<sup>5</sup>Examples provided in the online repository

as set membership constraints where sets may be clusters, or documents.

One improvement that could be made to the proposed system is to facilitate the discovery of relations that are not one-to-one. While we found many isolated examples of one-to-many relations expressed in the frames, a strictly symmetrical proportional analogy does not seem ideal for capturing one-to-many relations.

As outlined by (Turney, 2006) there are many applications of automating the construction and/or discovery of analogical relations. Some of these include relation classification, metaphor detection, word sense disambiguation, information extraction, question answering and thesaurus generation.

Analogical frames should also provide insight into the geometry of word embeddings and may provide an interesting way to measure their quality.

The most interesting application of the system is in the area of computational creativity with a human in the loop. For example, analogical frames could be chosen for their interestingness and then expanded.

## 6.1 Software and Online Repository

The software implementing the proposed system as a set of command line applications can be found in the online repository<sup>6</sup>. The software is implemented in portable C++11 and compiles on both Windows and Unix based systems without compiled dependencies. Example outputs of the system as well as parameter settings are provided in the online repository including the outputs created from embeddings trained on a range of corpora.

## 7 Future Work and Conclusions

Further empirical evaluation is required. The establishment of more suitable empirical benchmarks for assessing the effectiveness of open analogy discovery is important. The most interesting potential application of this work is in the combination of automated discovery of analogies and human judgment. There is also the possibility of establishing a more open-ended compute architecture that could search continuously for analogical frames in an online fashion.

<sup>6</sup> <https://github.com/ldevine/AFM>

## References

- Paul Bartha. 2016. Analogy and analogical reasoning. *The Stanford Encyclopedia of Philosophy*.
- William Correa Beltran, Hélène Jaudoin, and Olivier Pivert. 2015. A clustering-based approach to the mining of analogical proportions. In *Tools with Artificial Intelligence (ICTAI), 2015 IEEE 27th International Conference on*, pages 125–131. IEEE.
- James P Blevins. 2016. *Word and paradigm morphology*. Oxford University Press.
- James P Blevins, Farrell Ackerman, Robert Malouf, J Audring, and F Masini. 2017. Word and paradigm morphology.
- Rens Bod. 2009. From exemplar to grammar: A probabilistic analogy-based model of language learning. *Cognitive Science*, 33(5):752–793.
- Leonidas AA Doumas, John E Hummel, and Catherine M Sandhofer. 2008. A theory of the discovery and predication of relational concepts. *Psychological review*, 115(1):1.
- Aleksandr Drozd, Anna Gladkova, and Satoshi Matsumura. 2016. Word embeddings, analogies, and machine learning: Beyond king+man+woman=queen. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3519–3530.
- Rashel Fam and Yves Lepage. 2016. Morphological predictability of unseen words using computational analogy.
- Rashel Fam and Yves Lepage. 2017. A study of the saturation of analogical grids agnostically extracted from texts.
- Robert M French. 2002. The computational modeling of analogy-making. *Trends in cognitive Sciences*, 6(5):200–205.
- Dedre Gentner. 1983. Structure-mapping: A theoretical framework for analogy. *Cognitive science*, 7(2):155–170.
- Dedre Gentner. 2010. Bootstrapping the mind: Analogical processes and symbol systems. *Cognitive Science*, 34(5):752–775.
- Dedre Gentner and Kenneth D Forbus. 2011. Computational models of analogy. *Wiley interdisciplinary reviews: cognitive science*, 2(3):266–276.
- Huda Hakami, Kohei Hayashi, and Danushka Bollegala. 2017. An optimality proof for the pairdiff operator for representing relations between words. *arXiv preprint arXiv:1709.06673*.
- Rogers P Hall. 1989. Computational approaches to analogical reasoning: A comparative analysis. *Artificial intelligence*, 39(1):39–120.
- Keith J Holyoak and Paul Thagard. 1989. Analogical mapping by constraint satisfaction. *Cognitive science*, 13(3):295–355.
- John E Hummel and Keith J Holyoak. 1997. Distributed representations of structure: A theory of analogical access and mapping. *Psychological review*, 104(3):427.
- Johannes Jurgovsky, Michael Granitzer, and Christin Seifert. 2016. Evaluating memory efficiency and robustness of word embeddings. In *European Conference on Information Retrieval*, pages 200–211. Springer.
- Paul Kiparsky. 1992. Analogy. *International Encyclopedia of Linguistics*.
- Kenneth J Kurtz, Chun-Hui Miao, and Dedre Gentner. 2001. Learning by analogical bootstrapping. *The Journal of the Learning Sciences*, 10(4):417–446.
- Philippe Langlais. 2016. Efficient identification of formal analogies.
- Jean-François Lavallée and Philippe Langlais. 2009. Unsupervised morphological analysis by formal analogy. In *Workshop of the Cross-Language Evaluation Forum for European Languages*, pages 617–624. Springer.
- Yves Lepage. 2014. Analogies between binary images: Application to chinese characters. In *Computational Approaches to Analogical Reasoning: Current Trends*, pages 25–57. Springer.
- Yves Lepage and Etienne Denoual. 2005. Purest ever example-based machine translation: Detailed presentation and assessment. *Machine Translation*, 19(3-4):251–282.
- Omer Levy, Yoav Goldberg, and Israel Ramat-Gan. 2014. Linguistic regularities in sparse and explicit word representations. In *CoNLL*, pages 171–180.
- Tal Linzen. 2016. Issues in evaluating semantic spaces using word analogies. *arXiv preprint arXiv:1606.07736*.
- Pedro Meseguer and Carme Torras. 2001. Exploiting symmetries within constraint satisfaction search. *Artificial intelligence*, 129(1-2):133–163.
- Laurent Miclet, Sabri Bayoudh, and Arnaud Delhay. 2008. Analogical dissimilarity: definition, algorithms and two experiments in machine learning. *Journal of Artificial Intelligence Research*, 32:793–824.
- Laurent Miclet and Jacques Nicolas. 2015. From formal concepts to analogical complexes. In *CLA 2015*, page 12.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. Linguistic regularities in continuous space word representations. In *HLT-NAACL*, volume 13, pages 746–751.

Francesca Rossi, Peter Van Beek, and Toby Walsh. 2006. *Handbook of constraint programming*. Elsevier.

Royal Skousen. 1989. *Analogical modeling of language*. Springer Science & Business Media.

Nicolas Stroppa and François Yvon. 2005. An analogical learner for morphological analysis. In *Proceedings of the Ninth Conference on Computational Natural Language Learning*, pages 120–127. Association for Computational Linguistics.

Peter D Turney. 2006. Similarity of semantic relations. *Computational Linguistics*, 32(3):379–416.

Peter D Turney. 2013. Distributional semantics beyond words: Supervised learning of analogy and paraphrase. *arXiv preprint arXiv:1310.5042*.

Ekaterina Vylomova, Laura Rimell, Trevor Cohn, and Timothy Baldwin. 2015. Take and took, gaggle and goose, book and read: Evaluating the utility of vector differences for lexical relation learning. *arXiv preprint arXiv:1509.01692*.

Yating Zhang, Adam Jatowt, and Katsumi Tanaka. 2016. Towards understanding word embeddings: Automatically explaining similarity of terms. In *Big Data (Big Data), 2016 IEEE International Conference on*, pages 823–832. IEEE.

# Specifying Conceptual Models Using Restricted Natural Language

**Bayzid Ashik Hossain**  
Department of Computing  
Sydney, NSW  
Australia

[bayzid-ashik.hossain@mq.edu.au](mailto:bayzid-ashik.hossain@mq.edu.au)

**Rolf Schwitter**  
Department of Computing  
Sydney, NSW  
Australia

[rolf.schwitter@mq.edu.au](mailto:rolf.schwitter@mq.edu.au)

## Abstract

The key activity to design an information system is conceptual modelling which brings out and describes the general knowledge that is required to build a system. In this paper we propose a novel approach to conceptual modelling where the domain experts will be able to specify and construct a model using a restricted form of natural language. A restricted natural language is a subset of a natural language that has well-defined computational properties and therefore can be translated unambiguously into a formal notation. We will argue that a restricted natural language is suitable for writing precise and consistent specifications that lead to executable conceptual models. Using a restricted natural language will allow the domain experts to describe a scenario in the terminology of the application domain without the need to formally encode this scenario. The resulting textual specification can then be automatically translated into the language of the desired conceptual modelling framework.

## 1 Introduction

It is well-known that the quality of an information system application depends on its design. To guarantee accurateness, adaptability, productivity and clarity, information systems are best specified at the conceptual level using a language with names for individuals, concepts and relations that are easily understandable by domain experts (Bernus et al., 2013). Conceptual modelling is the most important part of requirements engineering and is the first phase towards designing an information system (Olivé, 2007). The conceptual design procedure generally includes data, process and be-

havioral perceptions, and the actual database management system (DBMS) that is used to implement the design of the information system (Bernus et al., 2013). The DBMS could be based on any of the available data models. Designing a database means constructing a formal model of the desired application domain which is often called the universe of discourse (UOD). Conceptual modelling involves different parties who sit together and define the UOD. The process of conceptual modelling starts with the collection of necessary information from the domain experts by the knowledge engineers. The knowledge engineers then use traditional modelling techniques to design the system based on the collected information.

To design the database, a clear understanding of the application domain and an unambiguous information representation scheme is necessary. Object role modelling (ORM) (Halpin, 2009) makes the database design process simple by using natural language for verbalization, as well as diagrams which can be populated with suitable examples and by adding the information in terms of simple facts. On the other hand, entity relationship modelling (ERM) (Richard, 1990; Frantiska, 2018) does this by considering the UOD in terms of entities, attributes and relationships. Object-oriented modelling techniques such as the unified modelling language (UML) (O'Regan, 2017) provide a wide variety of functionality for specifying a data model at an implementation level which is suitable for the detailed design of an object oriented system. UML can be used for database design in general because its class diagrams provide a comprehensive entity-relationship notation that can be annotated with database constructs.

Alternatively, a Restricted Natural Language (RNL) (Kuhn, 2014) can be used by the domain experts to specify system requirements for conceptual modelling. A RNL can be defined as a subset of a natural language that is acquired by

constraining the grammar and vocabulary in order to reduce or remove its ambiguity and complexity. These RNLs are also known as controlled natural language (CNL) (Schwitzer, 2010). RNLs fall into two categories: 1. those that improve the readability for human beings especially for non-native speakers, and 2. those that facilitate the automated translation into a formal target language. The main benefits of an RNL are: they are easy to understand by humans and easy to process by machines. In this paper, we show how an RNL can be used to write a specification for an information system and how this specification can be processed to generate a conceptual diagram. The grammar of our RNL specifies and restricts the form of the input sentences. The language processor translates RNL sentences into a version of description logic. Note that the conceptual modelling process usually starts from scratch and therefore cannot rely on existing data that would make this process immediately suitable for machine learning techniques.

## 2 Related Work

There has been a number of works on formalizing conceptual models for verification purposes (Berardi et al., 2005; Calvanese, 2013; Lutz, 2002). This verification process includes consistency and redundancy checking. These approaches first represent the domain of interest as a conceptual model and then formalize the conceptual model using a formal language. The formal representation can be used to reason about the domain of interest during the design phase and can also be used to extract information at run time through query answering.

Traditional conceptual modelling diagrams such as entity relationship diagrams and unified modelling language diagrams are easy to generate and easily understandable for the knowledge engineers. These modelling techniques are well established. The problems with these conventional modelling approaches are: they have no precise semantics and no verification support; they are not machine comprehensible and as a consequence automated reasoning on the conceptual diagrams is not possible. Previous approaches used logic to formally represent the diagrams and to overcome these problems. The description logic (DL) *ALCQI* is well suited to do reasoning with entity relationship diagrams (Lutz, 2002), UML

class diagrams (Berardi et al., 2005) and ORM diagrams (Franconi et al., 2012). The DL *ALCQI* is an extension of the basic propositionally closed description logic *AL* and includes complex concept negation, qualified number restriction, and inverse role.

Table 1 shows the constructs of the *ALCQI* description logic with suitable examples. It is reported that finite model reasoning with *ALCQI* is decidable and ExpTime-complete<sup>1</sup>. Using logic to formally represent the conceptual diagrams introduces some problems too. For example, it is difficult to generate logical representations, in particular for domain experts; it is also difficult for them to understand these representations and no well established methodologies are available to represent the conceptual models formally. A solution to these problems is to use a RNL for the specification of conceptual models. There exist several ontology editing and authoring tools such as AceWiki (Kuhn, 2008), CLOnE (Funk et al., 2007), RoundTrip Ontology Authoring (Davis et al., 2008), Rabbit (Denaux et al., 2009), Owl Simplified English (Power, 2012) that already use RNL for the specification of ontologies; they translate a specification into a formal notation. There are also works on mapping formal notation into conceptual models (Brockmans et al., 2004; Bagui, 2009).

## 3 Proposed Approach

Several approaches have been proposed to use logic with conventional modelling techniques to verify the models and to get the semantics of the domains. These approaches allow machines to understand the models and thus support automated reasoning. To overcome the disadvantages associated with these approaches, we propose to use an RNL as a language for specifying conceptual models. The benefits of an RNL are: 1. the language is easy to write and understand for domain experts as it is a subset of a natural language, 2. the language gets its semantics via translation into a formal notation, and 3. the resulting formal notation can be used further to generate conceptual models.

Unlike previous approaches, we propose to write a specification of the conceptual model in RNL first and then translate this specification into description logic. Existing description logic rea-

---

<sup>1</sup><http://www.cs.man.ac.uk/~ezolin/dl/>

Construct	Syntax	Example
atomic concept	$C$	Student
atomic role	$P$	hasChild
atomic negation	$\neg C$	$\neg$ Student
conjunction	$C \sqcap D$	Student $\sqcap$ Teacher
(unqual.) exist. restriction	$\exists R$	$\exists$ hasChild
universal value restriction	$\forall R.C$	$\forall$ hasChild.Male
full negation	$\neg(C \sqcap D)$	$\neg$ (Student $\sqcap$ Teacher)
qual. cardinality restrictions	$\geq nR.C$	$\geq 2$ hasChild.Female
inverse role	$p^-$	$\exists$ hasChild $^-$ .Teacher

Table 1: The DL *ALCQI*.

soners<sup>2,3</sup> can be used to check the consistency of the formal notation and after that desired conceptual models can be generated from this notation. Our approach is to derive the conceptual model from the specification whereas in conventional approaches knowledge engineers first draw the model and then use programs to translate the model into a formal notation (Fillottrani et al., 2012). Figure 1 shows the proposed system architecture for conceptual modelling.

### 3.1 Scenario

Let's consider an example scenario of a learning management system for a university stated below:

*A Learning Management System (LMS) keeps track of the units the students do during their undergraduate or graduate studies at a particular university. The university offers a number of programs and each program consists of a number of units. Each program has a program name and a program id. Each unit has a unit code and a unit name. A student can take a number of units whereas a unit has a number of students. A student must study at least one unit and at most four units. Every student can enrol into exactly one program. The system stores a student id and a student name for each student.*

We reconstruct this scenario in RNL and after that the language processor translates the RNL specification into description logic using a feature-based phrase structure grammar (Bird et al., 2009). Our RNL consists of function words and content words. Function words (e.g., determiners, quantifiers and operators) describe the structure of the

RNL and their number is fixed. Content words (e.g, nouns and verbs) are domain specific and can be added to the lexicon during the writing process. The reconstruction process of this scenario in RNL is supported by a look-ahead text editor (Guy and Schwitter, 2017). The reconstructed scenario in RNL looks as follows:

- (a) No student is a unit.
- (b) No student is a program.
- (c) Every student is enrolled in exactly one program.
- (d) Every student studies at least one unit and at most four units.
- (e) Every student has a student id and has a student name.
- (f) No program is a unit and is a student.
- (g) Every program is composed of a unit.
- (h) Every program is enrolled by a student.
- (i) Every program has a program id and has a program name.
- (j) No unit is a student and is a program.
- (k) Every unit is studied by a student.
- (l) Every unit belongs to a program.
- (m) Every unit has a unit code and has a unit name.

Additionally, we use the following terminological statements expressed in RNL:

- (n) The verb studies is the inverse of the verb studied by.

<sup>2</sup><https://franz.com/agraph/racer/>

<sup>3</sup><http://www.hermit-reasoner.com/>

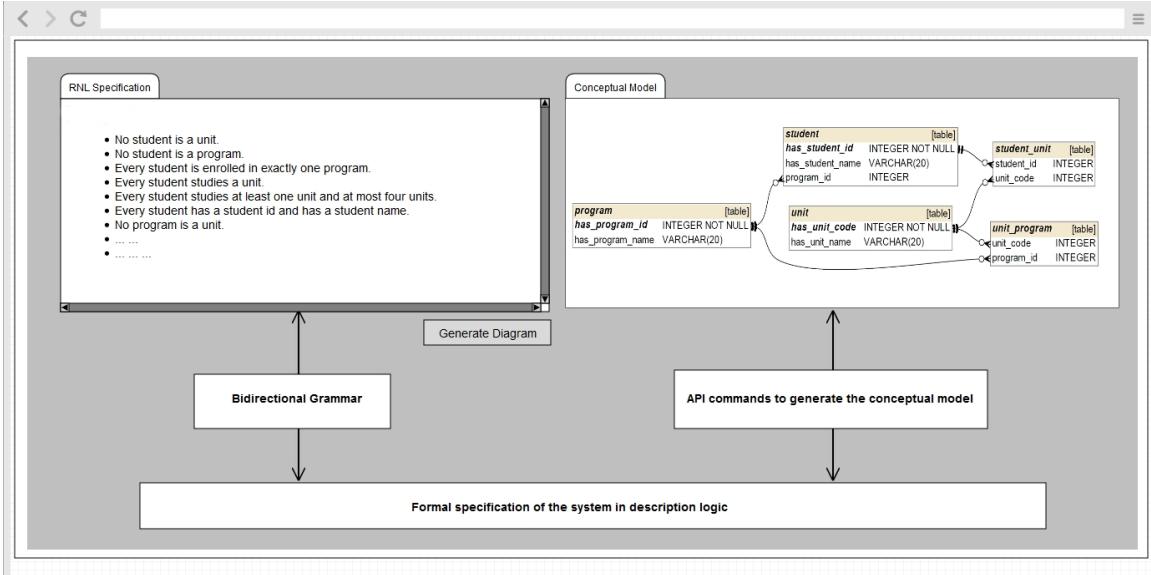


Figure 1: Proposed system architecture for conceptual modelling using restricted natural language.

- (o) The verb composed of is the inverse of the verb belongs to.

### 3.2 Grammar

A feature-based phrase structure grammar has been built using the NLTK (Loper and Bird, 2002) toolkit to parse the above-mentioned specification. The resulting parse trees for these sentences are then translated by the language processor into their equivalent description logic statements. Below we show a scaffolding of the grammar rules with feature structures that we used for our case study:

```

S ->
NP [NUM=?n, FNC=subj]
VP [NUM=?n]
FS

VP [NUM=?n] ->
V [NUM=?n] NP [FNC=obj] |
V [NUM=?n] Neg NP [NUM=?n, FNC=obj] |
VP [NUM=?n] CC VP [NUM=?n]

NP [NUM=?n, FNC=subj] ->
UQ [NUM=?n] N [NUM=?n] |
NQ [NUM=?n] N [NUM=?n] |
Det [NUM=?n] N [NUM=?n] |
KP [NUM=?n] VB [NUM=?n] |
KP [NUM=?n] VBN [NUM=?n]

NP [NUM=?n, FNC=obj] ->
Det [NUM=?n] N [NUM=?n] |
RB [NUM=?n] CD [NUM=?n] N [NUM=?n] |
KP [NUM=?n] VB [NUM=?n] |
KP [NUM=?n] VBN [NUM=?n]

V [NUM=?n] ->
Copula [NUM=?n] |
VB [NUM=?n] |
Copula [NUM=?n] VBN [NUM=?n] |

```

```

Copula [NUM=?n] JJ [NUM=?n]
VB [NUM=pl] -> "study" | ...
VB [NUM=sg] -> "studies" | ...
VBN -> "studied" "by" | ...
Copula [NUM=sg] -> "is"
Copula [NUM=pl] -> "are"

JJ -> "inverse" "of"
CC -> "and" | "or"

Det [NUM=sg] -> "A" | "a" | ...
Det -> "The" | "the"

UQ [NUM=sg] -> "Every"
NQ -> "No"

Neg -> "not"

N [NUM=sg] -> "student" | ...
N [NUM=pl] -> "students" | ...

RB -> "exactly" | ...

CD [NUM=sg] -> "one"
CD [NUM=pl] -> "two" | ... | "four"

KP -> "The" "verb" | ...

FS -> "."

```

In order to translate the resulting syntax trees into the description logic representation, we have used the owl/xml syntax<sup>4</sup> of Web Ontology Language (OWL) as the formal target notation.

### 3.3 Case Study

The translation process starts by reconstructing the specification in RNL that follows the rules of the

<sup>4</sup><https://www.w3.org/TR/owl-xmlsyntax/>

feature based grammar. While writing the specification in RNL, we tried to use the same vocabulary as in the natural language description.

The first two sentences of our specification use a negative quantifier in subject position and an indefinite determiner in object position:

(a) *No student is a unit.*

(b) *No student is a program.*

The translation of the sentence (a) into owl/xml notation results in the declaration of two atomic classes `student` and `unit` which are disjoint from each other.

```
<Declaration>
  <Class IRI="#student"/>
</Declaration>
<Declaration>
  <Class IRI="#unit"/>
</Declaration>
<DisjointClasses>
  <Class IRI="#student"/>
  <Class IRI="#unit"/>
</DisjointClasses>
```

Similarly, the translation of the sentence (b) results in the declaration of two disjoint atomic classes `student` and `program`. Both sentences (a) and (b) are related to expressing atomic negation in the DL *ALCQI* (see table 1).

```
<Declaration>
  <Class IRI="#student"/>
</Declaration>
<Declaration>
  <Class IRI="#program"/>
</Declaration>
<DisjointClasses>
  <Class IRI="#student"/>
  <Class IRI="#program"/>
</DisjointClasses>
```

The RNL that we have designed for this case study allows for verb phrase coordination; for example, the two above-mentioned sentences (a+b) can be combined in the following way:

(a+b) *No student is a unit and is a program.*

The translation of this sentence (a+b) results in the declaration of three atomic classes `student`, `unit` and `program` where `student` is disjoint from both `unit` and `program`.

```
<Declaration>
  <Class IRI="#student"/>
</Declaration>
<Declaration>
  <Class IRI="#unit"/>
</Declaration>
<Declaration>
  <Class IRI="#program"/>
```

```
</Declaration>
<DisjointClasses>
  <Class IRI="#student"/>
  <Class IRI="#unit"/>
</DisjointClasses>
<DisjointClasses>
  <Class IRI="#student"/>
  <Class IRI="#program"/>
</DisjointClasses>
```

Now let us consider the following RNL sentences that use a universal quantifier in subject position and a quantifying expression in object position:

(c) *Every student is enrolled in exactly one program.*

(d) *Every student studies at least one unit and at most four units.*

The universally quantified sentence (c) which contains a cardinality quantifier in the object position is translated into an object property `enrolled_in` that has the class `student` as domain and the class `program` as range with an exact cardinality of 1. This corresponds to a qualified cardinality restriction in the DL *ALCQI*.

```
<Declaration>
  <ObjectProperty IRI="#enrolled_in"/>
</Declaration>
<ObjectPropertyDomain>
  <ObjectProperty IRI="#enrolled_in"/>
  <Class IRI="#student"/>
</ObjectPropertyDomain>
<ObjectPropertyRange>
  <ObjectProperty IRI="#enrolled_in"/>
  <ObjectExactCardinality cardinality="1">
    <ObjectProperty IRI="#enrolled_in"/>
    <Class IRI="#program"/>
  </ObjectExactCardinality>
</ObjectPropertyRange>
```

The universally quantified sentence (d) which has a compound cardinality quantifier in object position is translated into the object property `study` that has the class `student` as domain and the class `unit` as range with a minimum cardinality of 1 and maximum cardinality of 4. The translation of this sentence corresponds to a qualified cardinality restriction in the DL *ALCQI*.

```
<Declaration>
  <ObjectProperty IRI="#study"/>
</Declaration>
<ObjectPropertyDomain>
  <ObjectProperty IRI="#study"/>
  <Class IRI="#student"/>
</ObjectPropertyDomain>
<ObjectPropertyRange>
  <ObjectProperty IRI="#study"/>
  <ObjectMinCardinality cardinality="1">
    <ObjectProperty IRI="#study"/>
  </ObjectMinCardinality>
</ObjectPropertyRange>
```

```

<Class IRI="#unit"/>
</ObjectMinCardinality>
</ObjectPropertyRange>
<ObjectPropertyRange>
  <ObjectProperty IRI="#study"/>
  <ObjectMaxCardinality cardinality="4">
    <ObjectProperty IRI="#study"/>
    <Class IRI="#unit"/>
  </ObjectMaxCardinality>
</ObjectPropertyRange>

```

The following RNL sentence has a universal quantifier in subject position and a coordinated verb phrase with indefinite noun phrases in object position:

- (e) *Every student has a student id and has a student name.*

The translation of this sentence (e) results in two data properties for the class `student`. The first data property is `has_student_id` with a data type `integer` and the second data property is `has_student_name` with the data type `string`:

```

<Declaration>
  <DataProperty IRI="\#has_student_id"/>
</Declaration>
<DataPropertyDomain>
  <DataProperty IRI="\#has_student_id"/>
  <Class IRI="#student"/>
</DataPropertyDomain>
<DataPropertyRange>
  <DataProperty IRI="\#has_student_id"/>
  <Datatype abbreviatedIRI="xsd:integer"/>
</DataPropertyRange>
<Declaration>
  <DataProperty IRI="\#has_student_name"/>
</Declaration>
<DataPropertyDomain>
  <DataProperty IRI="\#has_student_name"/>
  <Class IRI="#student"/>
</DataPropertyDomain>
<DataPropertyRange>
  <DataProperty IRI="\#has_student_name"/>
  <Datatype abbreviatedIRI="xsd:string"/>
</DataPropertyRange>

```

- (f) *Every program is composed of a unit.*

The universally quantified sentence (f) which contains an indefinite determiner in the object position is translated into the object property `composed_of` with the class `program` as domain and the class `unit` as range. This corresponds to an unqualified existential restriction in the DL *ALCQI*.

```

<Declaration>
  <ObjectProperty IRI="#composed_of"/>
</Declaration>
<ObjectPropertyDomain>
  <ObjectProperty IRI="#composed_of"/>

```

```

<Class IRI="#program"/>
</ObjectPropertyDomain>
<ObjectPropertyRange>
  <ObjectProperty IRI="#composed_of"/>
  <ObjectSomeValuesFrom>
    <ObjectProperty IRI="#composed_of"/>
    <Class IRI="#unit"/>
  </ObjectSomeValuesFrom>
</ObjectPropertyRange>

```

The following two RNL sentences have a definite determiner both in subject position and object position and specify lexical knowledge for the language processor:

- (g) *The verb studies is the inverse of the verb studied by.*
- (h) *The verb composed of is the inverse of the verb belongs to.*

The translation of these two sentences results in the specification of inverse object properties. The translation of sentence (g) leads to the object properties `study` and `studied_by` which are inverse object properties. Similarly, the translation of sentence (h) states that the object properties `composed_of` and `belong_to` are also inverse object properties. These statements correspond to the inverse role construct in the DL *ALCQI*.

```

<InverseObjectProperties>
  <ObjectProperty IRI="#study"/>
  <ObjectProperty IRI="#studied_by"/>
</InverseObjectProperties>
<InverseObjectProperties>
  <ObjectProperty IRI="#composed_of"/>
  <ObjectProperty IRI="#belong_to"/>
</InverseObjectProperties>

```

The rest of the specification is similar to the examples that we have discussed above.

## 4 Reasoning

After generating the owl/xml notation for the RNL specification, we use Owlready ([Lamy, 2017](#)) that includes the description logic reasoner Hermit ([Glimm et al., 2014](#)) for consistency checking. Owlready is a Python library for ontology-oriented programming that allows to load OWL 2.0 ontologies and performs various reasoning tasks. For example, consistency checking of the specification can be performed on the class level. If a domain expert writes for example "*No student is a unit*" and later specifies that "*Every unit is a student*", then the reasoner can detect this inconsistency and informs the domain expert about this conflict. The owl/xml notation below shows how

this inconsistency ("owl:*Nothing*") is reported after running the reasoner.

```
<rdf:Description
    rdf:about="http://www.w3.org/2002/07/owl#Nothing">
    <owl:equivalentClass
        rdf:resource=
            "http://www.w3.org/2002/07/owl#Nothing"/>
    <owl:equivalentClass rdf:resource="#student"/>
    <owl:equivalentClass rdf:resource="#unit"/>
</rdf:Description>
```

This inconsistency can be highlighted directly in the RNL specification; that means the domain expert can fix the textual specification and does not have to worry about the underlying formal notation.

## 5 Conceptual Model Generation

In the next step, we extract necessary information such as a list of entities (classes), attributes (data properties) and relationships (object properties) from the owl/xml file to generate the conceptual model. This information is extracted by executing XPath (Berglund et al., 2003)<sup>5</sup> queries over the owl/xml notation and then it is used to build a database schema containing a number of tables representing the entities with associated attributes. Relationships among the entities are represented by using foreign keys in the tables. An SQL script is generated containing SQLite commands<sup>6</sup> for this database schema.

This SQL script is executed by using SQLite to generate the corresponding database for the specification. After that, we use SchemaCrawler<sup>7</sup> to generate the entity relationship diagram (see fig. 2) from the SQL script. SchemaCrawler is a free database schema discovery and comprehension tool that allows to generate diagrams from SQL code.

For mapping a description logic representation to an entity relationship diagram, we have used the approach described by Algorithm 1. All the classes in the OWL file become entities in the ER-diagram. Object properties are mapped into relations between the entities and data properties are mapped into attributes for these entities. The qualified cardinality restrictions of the object properties define relationship cardinalities in the diagram.

We understand conceptual modelling as a round tripping process. That means a domain expert can

---

**Algorithm 1:** Mapping description logic representation to SQLite commands for generating entity relationship diagrams.

---

**Input:** Logical notation in description logic  
**Output:** SQLite script  
*entity\_list*= extract\_class(owl/xml file);  
*data\_property\_list*=  
  extract\_data\_property(owl/xml file);  
*object\_property\_list*=  
  extract\_object\_property(owl/xml file);  
**for** *entity* in *entity\_list* **do**  
  create\_table(*entity*)  
  **for** *data\_property* in  
    *data\_property\_list* **do**  
    add\_data\_property(*entity*,  
      *data\_property*)  
  **for** *object\_property* in  
    *object\_property\_list* **do**  
    **if** cardinality == 1 **then**  
      add\_data\_property(*entity*,  
        *data\_property*)  
    **end**  
  **end**  
**end**  
**for** *object\_property* in  
  *object\_property\_list* **do**  
  create\_relationship(*entity*,  
    *object\_property*)  
**end**

---

write the RNL specification first, then generate the conceptual model from the specification, and then a knowledge engineer might want to modify the conceptual model. These modifications will then be reflected on the level of the RNL by verbalising the formal notation. During this modification process the reasoner can be used to identify inconsistencies found in a given specification and to give appropriate feedback to the knowledge engineer on the graphical level or to the domain expert on the textual level.

## 6 Discussion

The outcome of our experiment justifies the proposed approach for conceptual modelling. We have used a phase structure grammar to convert a RNL specification into description logic. This experiment shows that it is possible to generate formal representations from RNL specifications and

<sup>5</sup>[https://www.w3schools.com/xml/xml\\_xpath.asp](https://www.w3schools.com/xml/xml_xpath.asp)

<sup>6</sup><https://www.sqlite.org/index.html>

<sup>7</sup><https://www.schemacrawler.com/>

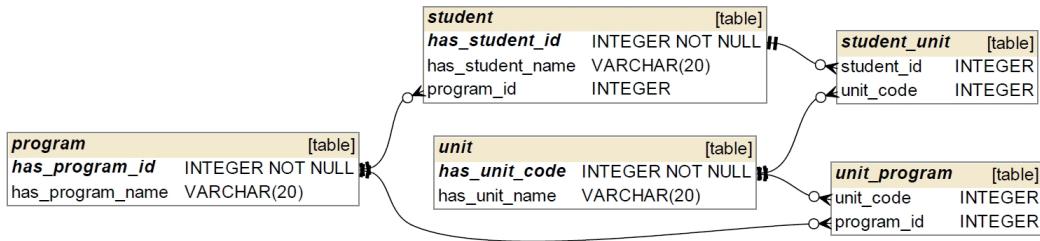


Figure 2: Entity relationship diagram generated from the formal representation using SchemaCrawler.

these formal representations can be mapped to different conceptual models. The proposed approach for conceptual modelling addresses two research challenges<sup>8</sup>: 1. providing the right set of modelling constructs at the right level of abstraction to enable successful communication among the stakeholders (i.e. domain experts, knowledge engineers, and application programmers); 2. preserving the ease of communication and enabling the generation of a database schema which is a part of the application software.

## 7 Future Work

We are planning to develop a fully-fledged restricted natural language for conceptual modelling of information systems. We want to use this language as a specification language that will help the domain experts to write the system requirements precisely. We are also planning to develop a conceptual modelling framework that will allow users to write specifications in RNL and will generate conceptual models from the specification. This tool will also facilitate the verbalization of the conceptual models and allow users to manipulate the models in a round tripping fashion (from specification to conceptual models and conceptual models to specifications). This approach has several advantages for the conceptual modelling process: Firstly, it will use a common formal representation to generate different conceptual models. Secondly, it will make the conceptual modelling process easy to understand by providing a framework to write specifications, generate visualizations, and verbalizations. Thirdly, it is machine-processable like other logical approaches and support verification; furthermore, verbalization will

facilitate better understanding of the modelling process which is only available in limited forms in the current conceptual modelling frameworks.

## 8 Conclusion

In this paper we demonstrated that an RNL can serve as a high-level specification language for conceptual modelling, in particular for specifying entity-relationship models. We described an experiment that shows how we can support the proposed modelling approach. We translated a specification of a conceptual model written in RNL into an executable description logic program that is used to generate the entity-relationship model. Our RNL is supported by automatic consistency checking, and is therefore very suitable for formalizing and verifying conceptual models. The presented approach is not limited to a particular modeling framework and can be used apart from entity-relationship models also for object-oriented models and object role models. Our approach has the potential to bridge the gap between a seemingly informal specification and a formal representation in the domain of conceptual modelling.

## References

- Sikha Bagui. 2009. Mapping OWL to the entity relationship and extended entity relationship models. *International Journal of Knowledge and Web Intelligence* 1(1-2):125–149.
- Daniela Berardi, Diego Calvanese, and Giuseppe De Giacomo. 2005. Reasoning on uml class diagrams. *Artificial Intelligence* 168(1):70–118.
- Anders Berglund, Scott Boag, Don Chamberlin, Mary F Fernández, Michael Kay, Jonathan Robbie, and Jérôme Siméon. 2003. **XML Path Language (XPath)**. World Wide Web Consortium (W3C) <https://www.w3.org/TR/xpath20/>.

<sup>8</sup><http://www.conceptualmodeling.org/ConceptualModeling.html>

- Peter Bernus, Kai Mertins, and Günter Schmidt. 2013. *Handbook on architectures of information systems*. Springer Science & Business Media.
- Steven Bird, Ewan Klein, and Edward Loper. 2009. Publications received. *Computational Linguistics* 36:283–284.
- Sara Brockmans, Raphael Volz, Andreas Eberhart, and Peter Löffer. 2004. Visual modeling of owl dl ontologies using uml. In *International Semantic Web Conference*. Springer, pages 198–213.
- Diego Calvanese. 2013. *Description Logics for Conceptual Modeling Forms of reasoning on UML Class Diagrams*. EPCL Basic Training Camp 2012-2013.
- Brian Davis, Ahmad Ali Iqbal, Adam Funk, Valentin Tablan, Kalina Bontcheva, Hamish Cunningham, and Siegfried Handschuh. 2008. Roundtrip ontology authoring. In *International Semantic Web Conference*. Springer, pages 50–65.
- Ronald Denaux, Vania Dimitrova, Anthony G Cohn, Catherine Dolbear, and Glen Hart. 2009. Rabbit to owl: ontology authoring with a cnl-based tool. In *International Workshop on Controlled Natural Language*. Springer, pages 246–264.
- Pablo R Fillottrani, Enrico Franconi, and Sergio Tesaris. 2012. The icom 3.0 intelligent conceptual modelling tool and methodology. *Semantic Web* 3(3):293–306.
- Enrico Franconi, Alessandro Mosca, and Dmitry Solomakhin. 2012. Orm2: formalisation and encoding in owl2. In *OTM Confederated International Conferences "On the Move to Meaningful Internet Systems"*. Springer, pages 368–378.
- Joseph Frantska. 2018. Entity-relationship diagrams. In *Visualization Tools for Learning Environment Development*, Springer, pages 21–30.
- Adam Funk, Valentin Tablan, Kalina Bontcheva, Hamish Cunningham, Brian Davis, and Siegfried Handschuh. 2007. Clone: Controlled language for ontology editing. In *The Semantic Web*, Springer, pages 142–155.
- Birte Glimm, Ian Horrocks, Boris Motik, Giorgos Stois, and Zhe Wang. 2014. HermiT: An OWL 2 Reasoner. *Journal of Automated Reasoning* 53(3):245–269.
- Stephen C Guy and Rolf Schwitter. 2017. The peng asp system: architecture, language and authoring tool. *Language Resources and Evaluation* 51(1):67–92.
- Terry Halpin. 2009. Object-role modeling. In *Encyclopedia of Database Systems*, Springer, pages 1941–1946.
- Tobias Kuhn. 2008. *Acewiki: A natural and expressive semantic wiki*. arXiv preprint [arXiv:0807.4618](https://arxiv.org/abs/0807.4618).
- Tobias Kuhn. 2014. A survey and classification of controlled natural languages. *Computational Linguistics* 40(1):121–170.
- Jean-Baptiste Lamy. 2017. Owlready: Ontology-oriented programming in python with automatic classification and high level constructs for biomedical ontologies. *Artificial intelligence in medicine* 80:11–28.
- Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics-Volume 1*. Association for Computational Linguistics, pages 63–70.
- Carsten Lutz. 2002. Reasoning about entity relationship diagrams with complex attribute dependencies. In *Proceedings of the International Workshop in Description Logics 2002 (DL2002)*, number 53 in CEUR-WS (<http://ceur-ws.org/>). pages 185–194.
- Gerard O’ Regan. 2017. Unified modelling language. In *Concise Guide to Software Engineering*, Springer, pages 225–238.
- Antoni Olivé. 2007. *Conceptual Modeling of Information Systems*. Springer-Verlag, Berlin, Heidelberg.
- Richard Power. 2012. Owl simplified english: a finite-state language for ontology editing. In *International Workshop on Controlled Natural Language*. Springer, pages 44–60.
- Barker Richard. 1990. *CASE Method: Entity Relationship Modelling*. Addison-Wesley Publishing Company, ORACLE Corporation UK Limited.
- Rolf Schwitter. 2010. Controlled natural languages for knowledge representation. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*. Association for Computational Linguistics, pages 1113–1121.

# Extracting structured data from invoices

Xavier Holt \*

Syph

xavier@syph.com

Andrew Chisholm \*

Syph

andy@syph.com

## Abstract

Business documents encode a wealth of information in a format tailored to human consumption – i.e. aesthetically dispersed natural language text, graphics and tables.

We address the task of extracting key fields (e.g. the amount due on an invoice) from a wide-variety of potentially unseen document formats. In contrast to traditional template driven extraction systems, we introduce a content-driven machine-learning approach which is both robust to noise and generalises to unseen document formats. In a comparison of our approach with alternative invoice extraction systems, we observe an absolute accuracy gain of 20% across compared fields, and a 25%–94% reduction in extraction latency.

## 1 Introduction

To unlock the potential of data in documents we must first interpret, extract and structure their content. For bills and invoices, data extraction enables a wide variety of downstream applications. Extraction of fields such as the amount due and biller information enable the automation of invoice payment for businesses. Moreover, extraction of information such as the daily usage or supply charge as found on an electricity bill (e.g. Figure 1) enables the aggregation of usage statistics over time and automated supplier switching advice. Manual annotation of document content is a time-consuming, costly and error-prone process (Klein et al., 2004). For many organisations, processing accounts payable or expense claims requires ongoing manual transcription for

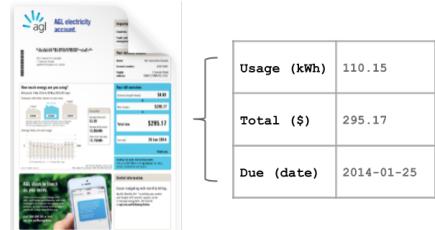


Figure 1: Energy bill with extracted fields.

verification of payment, supplier and pricing information. Template and RegEx driven extraction systems address this problem in part by shifting the burden of annotation from individual documents into the curation of extraction templates which cover a known document format. These approaches still necessitate ongoing human effort to produce reliable extraction templates as new supplier formats are observed and old formats change over time. This presents a significant challenge – Australia bill payments provider BPAY covers 26,000 different registered billers alone<sup>1</sup>.

We introduce SYPHT – a scaleable machine-learning solution to document field extraction. SYPHT combines OCR, heuristic filtering and a supervised ranking model conditioned on the content of document to make field-level predictions that are robust to variations in image quality, skew, orientation and content layout. We evaluate system performance on unseen document formats and compare 3 alternative invoice extraction systems on a common subset of key fields. Our system achieves the best results with an average accuracy of 92% across field types on unseen documents and the fastest median prediction latency of 3.8 seconds. We make our system available as an API<sup>2</sup> – enabling low latency key-field extraction scalable to hundreds of document per second.

<sup>1</sup>[www.bpay.com.au](http://www.bpay.com.au)

<sup>2</sup>[www.syph.com](http://www.syph.com)

## 2 Background

Information Extraction (IE) deals broadly with the problem of extracting structured information from unstructured text. In the domain of invoice and bill field extraction, document input is often better represented as a sparse arrangement of multiple text blocks rather than a single contiguous body of text. As financial documents are often themselves machine-generated, there is broad redundancy in this spatial layout of key fields across instances in a corpus. Early approaches exploit this structure by extracting known fields based on their relative position to extracted lines (Tang et al., 1995) and detected forms (Cesarini et al., 1998). Subsequent work aims to better generalise extractions patterns by constructing formal descriptions of document structure (Coüasnon, 2006) and developing systems which allow non-expert end-users to dynamically build extraction templates ad-hoc (Schuster et al., 2013). Similarly, the ITESOFT system (Rusiol et al., 2013) fits a term-position based extraction model from a small sample of human labeled samples which may be updated iteratively over time. More recently, D’Andecy et al. (2018) build upon this approach by incorporating an a-priori model of term-positions to their iterative layout-specific extraction model, significantly boosting performance on difficult fields.

While these approaches deliver high-precision extraction on observed document formats they cannot reliably or automatically generalise to unseen field layouts. Palm et al. (2017) present the closest work to our own with their CloudScan system for zero-shot field extraction from unseen invoice document forms. They train a recurrent neural network (RNN) model on a corpus of over 300K invoices to recognize 8 key fields, observing an aggregate F-score of 0.84 for fields extracted from held-out invoice layouts on their dataset. We consider a similar supervised approach but address the learning problem as one of value ranking in-place of sequence tagging. As they note, system comparison is complicated by a lack of a publicly available data for invoice extraction. Given the sensitive nature of invoices and prevalence of personally identifiable information, well-founded privacy concerns constrain open publishing in this domain. We address this limitation in part by rigorously anonymising a diverse set of invoices and submit them for evaluation to publicly available systems — without making public the data itself.

## 3 Task

We define the extraction task as follows: given a document and set of fields to query, provide the value of each field as it appears in the document. If there is no value for a given field present return `null`. This formulation is purely extractive – we do not consider implicit or inferred field values in our experiments or annotation. For example, while it may be possible to *infer* the value of tax paid with high confidence given the `net` and `gross` amount totals on an invoice, without this value being made explicit in text the correct system output is `null`. We do however consider inference over field names. Regardless of how a value is presented or labeled on a document, if it meets our query field definition systems must extract it. For example, valid `invoice number` values may be labeled as “*Reference*”, “*Document ID*” or even have no explicit label present. This canonicalization of field expression across document types is the core challenge addressed by extraction systems.

To compare system extractions we first normalise the surface form of extracted values by type. For example, dates expressed under a variety of formats are transformed to `yyyy-mm-dd` and numeric strings or reference number types (e.g. `ABN`, `invoice number`) have spaces and extraneous punctuation is removed. We adopt the evaluation scheme common to IE tasks such as Slot Filling (McNamee et al., 2009) and relation extraction (Mintz et al., 2009). For a given field predictions are judged *true-positive* if the predicted value matches the label; *false-positive* if the predicted value does not match the label; *true-negative* if both system and label are `null`; and *false-negative* if the predicted value is `null` and label is not `null`. In each instance we consider the type-specific normalised form for both value and label in comparisons. Standard metrics such as F-score or accuracy may then be applied to assess system performance.

Notably we do not consider the position of output values emitted by a system. In practise it is common to find multiple valid expressions of the same field at different points on a document – in this instance, labeling each value explicitly is both laborious for annotators and generally redundant. This may however incorrectly assign credit to systems for a missed predictions in rare cases, e.g. if both the `net` and `gross` totals normalise to the

same value (i.e. no applicable tax) a system may be marked correct for predicting either token for each field.

### 3.1 Fields

SYPHT provides extraction on a range of fields. For the scope of this paper and the sake of comparison, we restrict ourselves to the following fields relevant to invoices and bill payments:

**Supplier ABN** represents the Australian Business Number (ABN) of the invoice or bill supplier. For example, 16 627 246 039.

**Document Date** the date at which the document was released or printed. Generally distinct from the due date for bills and may be presented in a variety of formats, e.g. 11st December, 2018 or 11-12-2018.

**Invoice number** a reference generated by the supplier which uniquely identifies a document, e.g. INV-1447. Customer account numbers are not considered invoice references.

**Net amount** the total amount of new charges for goods and services, before taxes, discounts and other bill adjustments, e.g. \$50.00.

**GST** the amount of GST charged as it relates to the net amount of goods and services, e.g. \$5.00.

**Gross amount** the total gross cost of new charges for goods and services, including GST or any adjustments, e.g. \$55.00.

## 4 SYPHT

In this section we describe our end-to-end system for key-field extraction from business documents. We introduce a pipeline for field extraction at a high level and describe the prediction model and field annotation components in detail.

Although our system facilitates human-in-the-loop prediction validation, we do not utilise human-assisted predictions in our evaluation of system performance in Section 5.

**Preprocessing** documents are uploaded in a variety of formats (e.g. PDF or image files) and normalised to a common form of one-JPEG image per page. In development experiments we observe faster performance without degrading prediction accuracy by capping the rendered page resolution (~8MP) and limiting document colour channels to black and white.

**OCR** each page is independently parsed by an Optical Character Recognition (OCR) system in parallel which extracts textual tokens and their corresponding in-document positions.

**Filtering** for each query field we filter a subset of tokens as candidates in prediction based on the target field type. For example, we do not consider currency denominated values as candidate fills for a date field.

**Prediction** OCRed tokens and page images make up the input to our prediction model. For each field we rank the most likely value from the document for that field. If the most likely prediction falls below a tuned likelihood threshold, we emit `null` to indicate no field value is predicted. We describe our model implementation and training in Section 4.1.

**Validation** (optional) — consumers of the SYPHT API may specify a confidence threshold at which uncertain predictions are human validated before finalisation. We briefly describe our prediction assisted annotation and verification work-flow system in Section 4.2.

**Output** a JSON formatted object containing the extracted field-value pairs, model confidence and bounding-box information for each prediction is returned via an API call.

### 4.1 Model and training

Given an image and OCRed content as input, our model predicts the most likely value for a given query field. We use Spacy<sup>3</sup> to tokenise the OCR output. Each token is then represented through a wide range of features which describe the token’s syntactic, semantic, positional and visual content and context. We utilise part-of-speech tags, word-shape and other lexical features in conjunction with a sparse representation of the textual neighbourhood around a token to capture local textual context. In addition we capture a broad set of positional features including the  $x$  and  $y$  coordinates, in-document page offset and relative position of a token in relation to other predictions in the document. Our model additionally includes a range of proprietary engineered features tailored to field and document types of interest.

Field type information is incorporated into the model through token-level filtering. Examples of

---

<sup>3</sup>[spacy.io/models/en#en\\_core\\_web\\_sm](https://spacy.io/models/en#en_core_web_sm)

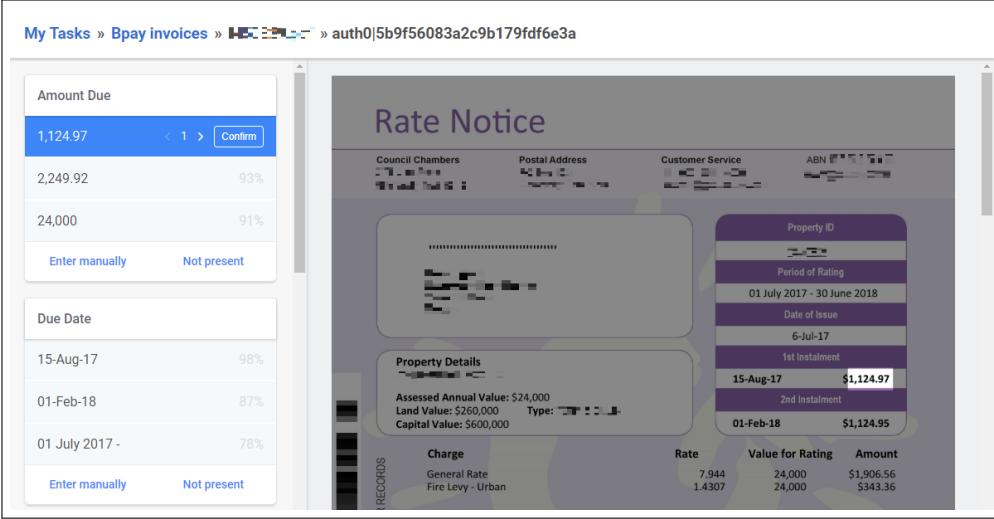


Figure 2: Our annotation and prediction verification tool — SYPHT VALIDATE. Tasks are presented with fields to annotate on the left and the source document for extraction on the right. We display the top predictions for each target field as suggestions for the user. In this example the most likely `Amount due` has been selected and the position of this prediction in the source document has been highlighted for confirmation.

field types which benefit from filtering are date, currency and integer fields; and fields with checksum rules. To handle multi-token field outputs, we utilise a combination of heuristic token merging (e.g. pattern based string combination for Supplier ABNs) and greedy token aggregation under a minimum sequence likelihood threshold from token level predictions (e.g. name and address fields).

We train our model by sampling instances at the token level. Matcher functions perform normalisation and comparison to annotated document labels for both for single and multi-token fields. All tokens which match the normalised form of the human-agreed value for a field are used to generate positive instances in a process analogous to distant supervision (Mintz et al., 2009). Other tokens in a document which match the field-type filter are randomly sampled as negative training instances. Instances of labels and sparse features are then used to train a gradient boosting decision tree model (LightGBM)<sup>4</sup>. To handle `null` predictions, we fit a threshold on token-level confidence which optimises a given performance metric; i.e. F-score for the models considered in this work. If the maximum likelihood value for a predicted token-sequence falls below the threshold for that field, a `null` prediction is returned instead.

## 4.2 Validation

An ongoing human annotation effort is often central to the training and evaluation of real-world machine learning systems. Well designed user-experiences for a given annotation task can significantly reduce the rate of manual-entry errors and speed up data collection (e.g. Prodigy<sup>5</sup>). We designed a predication-assisted annotation and validation tool for field extraction – SYPHT VALIDATE. Figure 2 shows a task during annotation.

Our tool is used to both supplement the training set and optionally – where field-level confidence does not meet a configurable threshold; provide human-in-the-loop prediction verification in real time. Suggestions are pre-populated through SYPHT predictions, transforming an otherwise tedious manual entry task into a relatively simple decision confirmation problem. Usability features such as token-highlighting and keyboard navigation greatly decrease the time it takes to annotate a given document.

We utilise continuous active learning by prioritising the annotation of new documents from our unlabeled corpus where the model is least confident. Conversely we observe high-confidence predictions which disagree with past human annotations are good candidates for re-annotation; often indicating the presence of annotation errors.

<sup>4</sup>[github.com/Microsoft/LightGBM](https://github.com/Microsoft/LightGBM)

<sup>5</sup><https://prodi.gy/>

### 4.3 Service architecture

SYPHT has been developed with performance at scale as a primary requirement. We use a micro-service architecture to ensure our system is both robust to stochastic outages and that we can scale up individual pipeline components to meet demand. Services interact via a dedicated message queue which increases fault-tolerance and ensure consistent throughput. Our system is capable of scaling to service a throughput of hundreds of requests per second at low latency to support mobile and other near real-time prediction use-cases. We consider latency a core metric for real-world system performance and include it in our evaluation of comparable systems in Section 5.

## 5 Evaluation

In this section we describe our methodology for creating the experimental dataset and system evaluation. We aim to understand how a variety of alternative extraction systems deals with various invoice formats. As a coarse representation of visual document structure, we compute a perceptual hash (Niu and Jiao, 2008) from the first-page of each document in a sample of Australian invoices. Personally identifiable information (PII) was then manually removed from each invoice by a human reviewer. SYPHT VALIDATE was used to generate the labels for the task, with between two and four annotators per field dependent on inter-annotator agreement. Annotators worked closely to ensure consistency between their labels and the data definitions listed in Section 3.1, with all fields having a sampled Cohen’s kappa greater than 0.8, and all fields except `net amount` having a kappa greater than 0.9. During the annotation procedure four documents were flagged as low quality and excluded from the evaluation set, resulting in a final count of 129. In each of these cases annotators could not reliably determine field values due to poor image quality. We evaluated against our deployed system after ensuring that all documents in the evaluation set were excluded from the model’s training set.

### 5.1 Compared systems

**ABBYY<sup>6</sup>** We ran ABBYY FlexiCapture 12 in batch mode on a modern quad-core desktop computer. While ABBYY software provides tools for creating extraction templates by hand, we utilised

<sup>6</sup>[www.abbyy.com/en-au/flexicapture/](http://www.abbyy.com/en-au/flexicapture/)

the generic invoice extraction model for parity with other comparison systems. By contrast with other systems which provided seamless API access, we operated the user interface manually and were unable to reliably record the distribution of prediction time per document. As such we only note the average extraction time aggregated over all test documents in Table 2

**EzzyBills**<sup>7</sup> automate data entry of invoice and account-payable in business accounting systems. We utilised the EzzyBills REST API.

**Rossum**<sup>8</sup> advertise a deep-learning driven data extraction API platform. We utilised their Python API<sup>9</sup> in our experiments.

## 6 Results

Table 1 presents accuracy results by field for each comparison system. SYPHT delivers the highest performance across measures fields with a macro averaged accuracy exceeding our comparable results by 23.7%, 22.8% and 20.2% (for Ezzy, ABBYY, Rossum respectively). Interestingly we observe low scores across the board on the `net amount` field with every systems performing significantly worse than the closely related `gross amount`. This field also obtained the lowest level of annotator agreement and was notoriously difficult to reliably assess – for example, the inclusion or exclusion of discounts, delivery costs and other adjustments to various sub totals on an invoice often complicates extraction.

The next best system Rossum performed surprising well considering their coverage of the European market; excluding support for Australian-specific invoice fields such as `ABN`. Still, even after excluding `ABN`, `net amount` and `GST` which may align to different field definitions, SYPHT maintains an 8 point accuracy advantage and more than 14 times lower median prediction latency.

Table 2 summarises the average prediction latency in seconds for each system alongside the times for documents at the 25th, 50th and 75th percentile of the response time distribution. Under the constraint of batch processing within the desktop ABBYY extraction environment we were unable to reliable record per-document prediction

<sup>7</sup>[www.ezzybills.com/api/](http://www.ezzybills.com/api/)

<sup>8</sup>[www.rossum.ai](http://www.rossum.ai)

<sup>9</sup>[pypi.org/project/rossum](http://pypi.org/project/rossum)

Field	Ezzy	ABBYY	Rossum	Ours
Supplier ABN	76.7	80.6	-	99.2
Invoice Number	72.1	82.2	86.8	94.6
Document Date	67.4	45.0	90.7	96.1
Net Amount	53.5	51.2	55.8	80.6
GST Amount	69.8	72.1	45.0	90.7
Gross Amount	75.2	89.1	84.5	95.3
Avg.	69.1	70.0	72.6	92.8

Table 1: Prediction accuracy by field.

	Avg.	25th	50th	75th
Rossum	67.06	47.7	54.4	91.0
Ezzy	27.9	20.6	26.9	34.5
ABBYY	5.6	-	-	-
Ours	4.2	3.3	3.8	4.8

Table 2: Prediction latency in seconds.

times and thus do not indicate their prediction response percentiles. SYPHT was faster than all comparison systems, and significantly faster relative to the other SaaS based API services. Even with the lack of network overhead inherent to ABBYY’s local extraction software, SYPHT maintains a 25% lower average prediction latency. In a direct comparison with other API based products we demonstrate stronger results still, with EzzyBills and Rossum being slower than SYPHT by a factor of 6.6 and 15.9 respectively in terms of mean prediction time per document.

## 7 Discussion and future work

While it is not a primary component of our current system, we have developed and continue to develop a number of solutions based on neural network models. Models for sequence labelling, such as LSTM (Gers et al., 1999) or Transformer (Vaswani et al., 2017) networks can be directly ensembled into the current system. We are also exploring the use of object classification and detection models to make use of the visual component of document data. Highly performant models such as YOLO (Redmon and Farhadi, 2018), are particularly interesting due to their ability to be used in real-time. We expect sub-5 second response times to constitute a rough threshold for realistic deployment of extraction systems in real time applications, making SYPHT the best system in contrast to either of the other two API-based services.

We also see an exciting opportunity to provide self-service model development – the ability for a customer to use their own documents to generate a model tailored to their set of fields. This would allow us to offer SYPHT for use cases where either we cannot or would not collect the prerequisite data. SYPHT VALIDATE provides a straightforward method for bootstrapping extraction models by providing rapid data annotation and efficient use of annotator time through active learning.

## 8 Conclusion

We present SYPHT, a SaaS API for key-field extraction from business documents. Our comparison with alternative extraction systems demonstrate both high accuracy and lower latency across extracted fields – enabling applications in real time for invoices and bill payment.

## Acknowledgements

We would like to thank members of the SYPHT team for their contributions to the system, annotation and evaluation effort: Duane Allam, Farzan Maghami, Paarth Arora, Raya Saeed, Saskia Parker, Simon Mittag and Warren Billington.

## References

- Francesca Cesarini, Marco Gori, Simone Marinai, and Giovanni Soda. 1998. Informys: A flexible invoice-like form-reader system. *IEEE Trans. Pattern Anal. Mach. Intell.* 20(7):730–745.
- Bertrand Coüasnon. 2006. Dmos, a generic document recognition method: application to table structure analysis in a general and in a specific way. *International Journal of Document Analysis and Recognition (IJDAR)* 8(2):111–122. <https://doi.org/10.1007/s10032-005-0148-5>.

- Vincent Poulain D'Andecy, Emmanuel Hartmann, and Marçal Rusiñol. 2018. Field extraction by hybrid incremental and a-priori structural templates. In *13th IAPR International Workshop on Document Analysis Systems, DAS 2018, Vienna, Austria, April 24-27, 2018*. pages 251–256. <https://doi.org/10.1109/DAS.2018.29>.
- Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. 1999. Learning to forget: Continual prediction with lstm .
- Bertin Klein, Stevan Agne, and Andreas Dengel. 2004. Results of a study on invoice-reading systems in germany. In Simone Marinai and Andreas R. Dengel, editors, *Document Analysis Systems VI*. Springer Berlin Heidelberg, Berlin, Heidelberg, pages 451–462.
- Paul McNamee, Heather Simpson, and Hoa Trang Dang. 2009. Overview of the TAC 2009 Knowledge Base Population Track. In *Proceedings of the 2009 Text Analysis Conference*.
- Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*. Association for Computational Linguistics, Stroudsburg, PA, USA, volume 2, pages 1003–1011. <http://dl.acm.org/citation.cfm?id=1690219.1690287>.
- Xia-mu Niu and Yu-hua Jiao. 2008. An overview of perceptual hashing. *Acta Electronica Sinica* 36(7):1405–1411.
- Rasmus Berg Palm, Ole Winther, and Florian Laws. 2017. Cloudscan - A configuration-free invoice analysis system using recurrent neural networks. In *14th IAPR International Conference on Document Analysis and Recognition, ICDAR 2017, Kyoto, Japan, November 9-15, 2017*. pages 406–413. <https://doi.org/10.1109/ICDAR.2017.74>.
- Joseph Redmon and Ali Farhadi. 2018. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767* .
- Maral Rusiol, Tayeb Benkhelfallah, and Vincent Poulain D'Andecy. 2013. Field extraction from administrative documents by incremental structural templates. In *Proceedings of the 12th International Conference on Document Analysis and Recognition*. IEEE Computer Society, pages 1100–1104.
- Daniel Schuster, Klemens Muthmann, Daniel Esser, Alexander Schill, Michael Berger, Christoph Weidling, Kamil Aliyev, and Andreas Hofmeier. 2013. Intellix – end-user trained information extraction for document archiving. In *Proceedings of the 2013 12th International Conference on Document Analysis and Recognition*. IEEE Computer Society, Washington, DC, USA, ICDAR '13, pages 101–105. <https://doi.org/10.1109/ICDAR.2013.28>.
- Y. Y. Tang, C. Y. Suen, Chang De Yan, and M. Cheriet. 1995. Financial document processing based on staff line and description language. *IEEE Transactions on Systems, Man, and Cybernetics* 25(5):738–754. <https://doi.org/10.1109/21.376488>.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*. pages 5998–6008.

## **Short papers**

# Exploring Textual and Speech information in Dialogue Act Classification with Speaker Domain Adaptation

Xuanli He\*

Monash University

xuanli.he1@monash.edu

Quan Hung Tran\*

Monash University

hung.tran@monash.edu

William Havard

Univ. Grenoble Alpes

william.havard@gmail.com

Laurent Besacier

Univ. Grenoble Alpes

laurent.besacier@imag.fr

Ingrid Zukerman

Monash University

ingrid.zukerman@monash.edu

Gholamreza Haffari

Monash University

gholamreza.haffari@monash.edu

## Abstract

In spite of the recent success of Dialogue Act (DA) classification, the majority of prior works focus on text-based classification with oracle transcriptions, i.e. human transcriptions, instead of Automatic Speech Recognition (ASR)'s transcriptions. Moreover, the performance of this classification task, because of speaker domain shift, may deteriorate. In this paper, we explore the effectiveness of using both acoustic and textual signals, either oracle or ASR transcriptions, and investigate speaker domain adaptation for DA classification. Our multimodal model proves to be superior to the unimodal models, particularly when the oracle transcriptions are not available. We also propose an effective method for speaker domain adaptation, which achieves competitive results.

## 1 Introduction

Dialogue Act (DA) classification is a sequence-labelling task, mapping a sequence of utterances to their corresponding DAs. Since DA classification plays an important role in understanding spontaneous dialogue (Stolcke et al., 2000), numerous techniques have been proposed to capture the semantic correlation between utterances and DAs.

Earlier on, statistical techniques such as Hidden Markov Models (HMMs) were widely used to recognise DAs (Stolcke et al., 2000; Julia et al., 2010). Recently, due to the enormous success of neural networks in sequence labeling/transduction tasks (Sutskever et al., 2014; Bahdanau et al., 2014; Popov, 2016), several recurrent neural network (RNN) based architectures have been proposed to conduct DA classification, resulting in

promising outcomes (Ji et al., 2016; Shen and Lee, 2016; Tran et al., 2017a).

Despite the success of previous work in DA classification, there are still several fundamental issues. Firstly, most of the previous works rely on transcriptions (Ji et al., 2016; Shen and Lee, 2016; Tran et al., 2017a). Fewer of these focus on combining speech and textual signals (Julia et al., 2010), and even then, the textual signals in these works utilise the oracle transcriptions. We argue that in the context of a spoken dialog system, oracle transcriptions of utterances are usually not available, i.e. the agent does not have access to the human transcriptions. Speech and textual data complement each other, especially when textual data is from ASR systems rather than oracle transcripts. Furthermore, domain adaptation in text or speech-based DA classification is relatively under-investigated. As shown in our experiments, DA classification models perform much worse when they are applied to new speakers.

In this paper, we explore the effectiveness of using both acoustic and textual signals, and investigate speaker domain adaptation for DA classification. We present a multimodal model to combine text and speech signals, which proves to be superior to the unimodal models, particularly when the oracle transcriptions are not available. Moreover, we propose an effective unsupervised method for speaker domain adaptation, which learns a suitable encoder for the new domain giving rise to representations similar to those in the source domain.

## 2 Model Description

In this section, we describe the basic structure of our model, which combines the textual and speech modalities. We also introduce a representation learning approach using adversarial ideas to tackle the domain adaptation problem.

\*Equal contribution

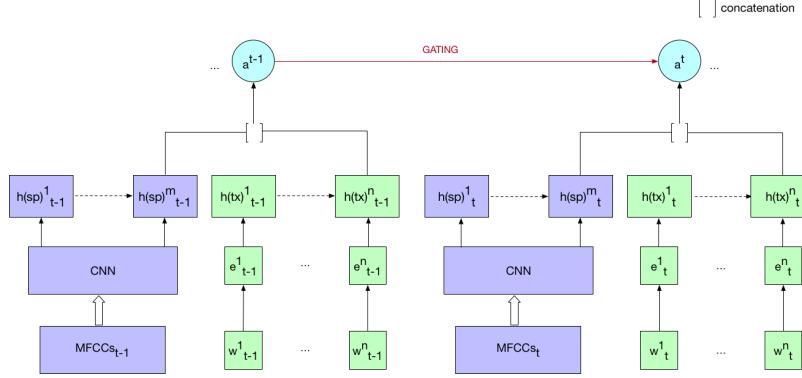


Figure 1: The multimodal model. For the utterance  $t$ , the left and right sides are encoded speech and text, respectively.

## 2.1 Our Multimodal Model

A conversation is comprised of a sequence of utterances  $\mathbf{u}_1, \dots, \mathbf{u}_T$ , and each utterance  $\mathbf{u}_t$  is labeled with a DA  $a_t$ . An utterance could include text, speech or both. We focus on online DA classification, and our classification model attempts to directly model the conditional probability  $p(\mathbf{a}_{1:T}|\mathbf{u}_{1:T})$  decomposed as follows:

$$p(\mathbf{a}_{1:T}|\mathbf{u}_{1:T}) = \prod_{t=1}^T p(a_t|a_{t-1}, \mathbf{u}_t). \quad (1)$$

According to Eqn. 1, during the training time the previous label is from the ground-truth data, while this information comes from the model during the inference stage. This discrepancy, referred as *label bias*, can result in error accumulation. To incorporate the previous DA information and mitigate the label-bias problem, we adopt the uncertainty propagation architecture (Tran et al., 2017b). The conditional probability term in Eqn. 1 is computed as follows:

$$\begin{aligned} a_t | a_{t-1}, \mathbf{u}_t &\sim \mathbf{q}_t \\ \mathbf{q}_t &= \text{softmax}(\bar{\mathbf{W}} \cdot \mathbf{c}(\mathbf{u}_t) + \bar{\mathbf{b}}) \\ \bar{\mathbf{W}} &= \sum_a \mathbf{q}_{t-1}(a) \mathbf{W}^a, \quad \bar{\mathbf{b}} = \sum_a \mathbf{q}_{t-1}(a) \mathbf{b}^a \end{aligned}$$

where  $\mathbf{W}^a$  and  $\mathbf{b}^a$  are DA-specific parameters gated on the DA  $a$ ,  $\mathbf{c}(\mathbf{u}_t)$  is the encoding of the utterance  $\mathbf{u}_t$ , and  $\mathbf{q}_{t-1}$  represents the uncertainty distribution over the DAs at the time step  $t - 1$ .

**Text Utterance.** An utterance  $\mathbf{u}_t$  includes a list of words  $w_t^1, \dots, w_t^n$ . The word  $w_t^i$  is embedded by  $\mathbf{x}_t^i = e(w_t^i)$  where  $e$  is an embedding table.

**Speech Utterance.** We apply a frequency-based transformation on raw speech signals to acquire

Mel-frequency cepstral coefficients (MFCCs), which have been very effective in speech recognition (Mohamed et al., 2012). To learn the context-specific features of the speech signal, a convolutional neural network (CNN) is employed over MFCCs:

$$\mathbf{x}'_t^1, \dots, \mathbf{x}'_t^m = \text{CNN}(\mathbf{s}_t^1, \dots, \mathbf{s}_t^k)$$

where  $\mathbf{s}_t^i$  is a MFCC feature vector at the position  $i$  for the  $t$ -th utterance.

**Encoding of Text+Speech.** As illustrated in Figure 1, we employ two RNNs with LSTM units to encode the text and speech sequences of an utterance  $\mathbf{u}_t$ :

$$\begin{aligned} \mathbf{c}(\mathbf{u}_t)^{tx} &= \text{RNN}_{\theta}(\mathbf{x}_t^1, \dots, \mathbf{x}_t^n) \\ \mathbf{c}(\mathbf{u}_t)^{sp} &= \text{RNN}_{\theta'}(\mathbf{x}'_t^1, \dots, \mathbf{x}'_t^m). \end{aligned}$$

where the encoding of the text  $\mathbf{c}(\mathbf{u}_t)^{tx}$  and speech  $\mathbf{c}(\mathbf{u}_t)^{sp}$  are the last hidden states of the corresponding RNNs whose parameters are denoted by  $\theta$  and  $\theta'$ . The distributed representation  $\mathbf{c}(\mathbf{u}_t)$  of the utterance  $\mathbf{u}_t$  is then the concatenation of  $\mathbf{c}(\mathbf{u}_t)^{tx}$  and  $\mathbf{c}(\mathbf{u}_t)^{sp}$ .

## 2.2 Speaker Domain Adaptation

Different people tend to speak differently. This creates a problem for DA classification systems, as unfamiliar speech signals might not be recognised properly. In our preliminary experiments, the performance of DA classification on speakers that are unseen in the training set suffers from dramatic performance degradation over test set. This motivates us to explore the problem of speaker domain adaptation in DA classification.

We assume we have a large amount of labelled source data pair  $\{X_{src}, Y_{src}\}$ , and a small amount

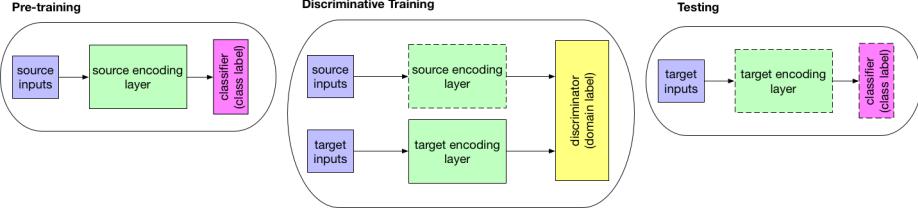


Figure 2: Overview of discriminative model. Dashed lines indicate frozen parts

of unlabelled target data  $X_{trg}$ , where an utterance  $\mathbf{u} \in X$  includes both speech and text parts. Inspired by Tzeng et al. (2017), our goal is to learn a target domain encoder which can fool a domain classifier  $C_\phi$  in distinguishing whether the utterance belongs to the source or target domain. Once the target encoder is trained to produce representations which look like those coming from the source domain, the target encoder can be used together with other components of the source DA prediction model to predict DAs for the target domain (see Figure 2).

We use a 1-layer feed-forward network as the domain classifier:

$$C_\phi(\mathbf{r}) = \sigma(\mathbf{W}_C \cdot \mathbf{r} + b_C)$$

where the classifier produces the probability of the input representation  $\mathbf{r}$  belonging to the source domain, and  $\phi$  denotes the classifier parameters  $\{\mathbf{W}_C, b_C\}$ . Let the target and source domain encoders are denoted by  $\mathbf{c}_{trg}(\mathbf{u}_{trg})$  and  $\mathbf{c}_{src}(\mathbf{u}_{trg})$ , respectively. The training objective of the domain classifier is:

$$\begin{aligned} \min_{\phi} \quad & \mathcal{L}_1(X_{src}, X_{trg}, C_\phi) = \\ & -\mathbb{E}_{\mathbf{u} \sim X_{src}} [\log C_\phi(\mathbf{c}_{src}(\mathbf{u}))] \\ & -\mathbb{E}_{\mathbf{u} \sim X_{trg}} [1 - \log C_\phi(\mathbf{c}_{trg}(\mathbf{u}))]. \end{aligned}$$

As mentioned before, we keep the source encoder fixed and train the parameters of the target domain encoder. The training objective of the target domain encoder is

$$\begin{aligned} \min_{\theta'_{trg}} \quad & \mathcal{L}_2(X_{trg}, C_\phi) = \\ & -\mathbb{E}_{\mathbf{u} \sim X_{trg}} [\log C_\phi(\mathbf{c}_{trg}(\mathbf{u}))] \end{aligned}$$

where the optimisation is performed over the speech RNN parameters  $\theta'_{trg}$  of the target encoder. We also tried to optimise other parameters (i.e. CNN parameters, word embeddings and text RNN parameters), but the performance is similar to the

speech RNN only. This is possibly because the major difference between source and target domain data is due to the speech signals. We alternate between optimising  $\mathcal{L}_1$  and  $\mathcal{L}_2$  by using Adam (Kingma and Ba, 2014) until a training condition is met.

### 3 Experiments

#### 3.1 Datasets

We test our models on two datasets: the MapTask Dialog Act data (Anderson et al., 1991) and the Switchboard Dialogue Act data (Jurafsky et al., 1997).

**MapTask dataset** This dataset consist of 128 conversations labelled with 13 DAs. We randomly partition this data into 80% training, 10% development and 10% test sets, having 103, 12 and 13 conversations respectively.

**Switchboard dataset** There are 1155 transcriptions of telephone conversations in this dataset, and each utterance falls into one of 42 DAs. We follow the setup proposed by Stolcke et al. (2000): 1115 conversations for training, 21 for development and 19 for testing. Since we do not have access to the original recordings of Switchboard dataset, we use synthetic speeches generated by a text-to-speech (TTS) system from the oracle transcriptions.

#### 3.2 Results

**In-Domain Evaluation.** Unlike most prior work (Ji et al., 2016; Shen and Lee, 2016; Tran et al., 2017a), we use ASR transcripts, produced by the CMUSphinx ASR system, rather than the oracle text. We argue that most dialogues in the real world are in the speech format, thus our setup is closer to the real-life scenario.

As shown in Tables 1 and 2, our multimodal model outperforms strong baselines on Switchboard and MapTask datasets, when using the ASR transcriptions. When using the oracle text, the in-

formation from the speech signal does not lead to further improvement though, possibly due to the existence of acoustic features (such as tones, question markers etc) in the high quality transcriptions. On MapTask, there is a large gap between oracle-based and ASR-based models. This degradation is mainly caused by the poor quality acoustic signals in MapTask, making ASR ineffective compared to directly predicting DAs from the speech signal.

Models	Accuracy
Oracle text	
Stolcke et al. (2000)	71.00%
Shen and Lee (2016)	72.60%
Tran et al. (2017a)	74.50%
Text only (ours)	74.97%
Text+Speech (ours)	74.98%
Speech and ASR	
Speech only	59.71%
Text only (ASR)	66.39%
Text+Speech (ASR)	<b>68.25%</b>

Table 1: Results of different models on Switchboard data.

Models	Accuracy
Oracle text	
Julia et al. (2010)	55.40%
Tran et al. (2017a)	61.60%
Text only (ours)	<b>61.73%</b>
Text+Speech (ours)	61.67%
Speech and ASR	
Speech only	39.32%
Text only (ASR)	38.10%
Text+Speech (ASR)	<b>39.39%</b>

Table 2: Results of different models on MapTask data.

**Out-of-Domain Evaluation.** We evaluate our domain adaptation model on the out of domain data on Switchboard. Our training data comprises of five known speakers, whereas development and test sets include data from three new speakers. The speeches for these 8 speakers are generated by a TTS system.

As described in Section 2.2, we pre-train our speech models on the labeled training data from the 5 known speakers, then train speech encoders

for the new speakers using speeches from both known and new speakers. During domain adaptation, the five known speakers are marked as the source domain, while the three new speakers are treated as the target domains. For domain adaptation with unlabelled data, the DA tags of both the source and target domains are removed. We test the source-only model and the domain adaptation models merely on the three new speakers in test data. As shown in Table 3, compared with the source-only model, the domain adaptation strategy improves the performance of speech-only and text+speech models, consistently and substantially.

Methods	Speech	Text+Speech
Unadapted	48.73%	63.57%
Domain Adapted	54.37%	67.21%
Supervised Learning	56.19 %	68.04%

Table 3: Experimental results of the unadapted (i.e. source-only) and domain adapted models using unlabeled data on Switchboard, as well as the supervised learning upperbound.

To assess the effectiveness of our domain adaptation architecture, we compare it with the supervised learning scenario where the model has access to labeled data from all speakers during training. To do this, we randomly add two thirds of labelled development data of new speakers to the training set, and apply the trained model to the test set. The supervised learning scenario is an upperbound to our domain adaptation approach, as it makes use of labeled data; see the results in the last row of Table 3. However, the gap between supervised learning and domain adaptation is not big compared to that between the adapted and unadapted models, showing that our domain adaptation technique has been effective.

## 4 Conclusion

In this paper, we have proposed a multimodal model to combine textual and acoustic signals for DA prediction. We have demonstrated that the our model exceeds unimodal models, especially when oracle transcriptions do not exist. In addition, we have proposed an effective domain adaptation technique in order to adapt our multimodal DA prediction model to new speakers.

## References

- Anne H Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, et al. 1991. The hcrc map task corpus. *Language and speech* 34(4):351–366.
- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR* abs/1409.0473.
- John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*. IEEE, volume 1, pages 517–520.
- Yangfeng Ji, Gholamreza Haffari, and Jacob Eisenstein. 2016. A latent variable recurrent neural network for discourse relation language models. *CoRR* abs/1603.01913.
- Fatema N. Julia, Khan M. Iftekharuddin, and Atiq U. Islam. 2010. Dialog act classification using acoustic and discourse information of maptask data. *International Journal of Computational Intelligence and Applications* 09(04):289–311.
- D. Jurafsky, E. Shriberg, and D. Biasca. 1997. Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual. Technical Report Draft 13, University of Colorado, Institute of Cognitive Science.
- Diederik P. Kingma and Jimmy Ba. 2014. *Adam: A method for stochastic optimization*. *CoRR* abs/1412.6980. <http://arxiv.org/abs/1412.6980>.
- Abdel-rahman Mohamed, Geoffrey Hinton, and Gerald Penn. 2012. Understanding how deep belief networks perform acoustic modelling. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, pages 4273–4276.
- Ramesh Nallapati, Bing Xiang, and Bowen Zhou. 2016. Sequence-to-sequence rnns for text summarization. *CoRR* abs/1602.06023.
- Alexander Popov. 2016. *Deep Learning Architecture for Part-of-Speech Tagging with Word and Suffix Embeddings*, Springer International Publishing, Cham, pages 68–77.
- Sheng-syun Shen and Hung-yi Lee. 2016. Neural attention models for sequence classification: Analysis and application to key term extraction and dialogue act detection. *CoRR* abs/1604.00077.
- Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca A. Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *CoRR* cs.CL/0006023.
- Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. *CoRR* abs/1409.3215.
- Quan Hung Tran, Ingrid Zukerman, and Gholamreza Haffari. 2017a. A hierarchical neural model for learning sequences of dialogue acts. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. volume 1, pages 428–437.
- Quan Hung Tran, Ingrid Zukerman, and Gholamreza Haffari. 2017b. Preserving distributional information in dialogue act classification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. pages 2141–2146.
- Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. 2017. Adversarial discriminative domain adaptation. *CoRR* abs/1702.05464.

# Cluster Labeling by Word Embeddings and WordNet’s Hypernymy

Hanieh Poostchi

University of Technology Sydney  
Capital Markets CRC  
hpoostchi@cmcrc.com

Massimo Piccardi

University of Technology Sydney  
massimo.piccardi@uts.edu.au

## Abstract

Cluster labeling is the assignment of representative labels to clusters of documents or words. Once assigned, the labels can play an important role in applications such as navigation, search and document classification. However, finding appropriately descriptive labels is still a challenging task. In this paper, we propose various approaches for assigning labels to word clusters by leveraging word embeddings and the synonymy and hypernymy relations in the WordNet lexical ontology. Experiments carried out using the WebAP document dataset have shown that one of the approaches stand out in the comparison and is capable of selecting labels that are reasonably aligned with those chosen by a pool of four human annotators.

## 1 Introduction and Related Work

Document collections are often organized into clusters of either documents or words to facilitate applications such as navigation, search and classification. The organization can prove more useful if its clusters are characterized by *sets of representative labels*. The task of assigning a set of labels to each individual cluster in a document organization is known as cluster labeling (Wang et al., 2014) and it can provide a useful description of the collection in addition to fundamental support for navigation and search.

In Manning et al. (2008), cluster labeling approaches have been subdivided into *i*) differential cluster labeling and *ii*) cluster-internal labeling. The former selects cluster labels by comparing the distribution of terms in one cluster with those of the other clusters while the latter selects labels that are solely based on each cluster indi-

vidually. Cluster-internal labeling approaches include computing the clusters’ centroids and using them as labels, or using lists of terms with highest frequencies in the clusters. However, all these approaches can only select cluster labels from the terms and phrases that explicitly appear in the documents, possibly failing to provide an appropriate level of abstraction or description (Lau et al., 2011). As an example, a word cluster containing words *dog* and *wolf* should not be labeled with either word, but as *canids*. For this reason, in this paper we explore several approaches for labeling word clusters obtained from a document collection by leveraging the synonymy and hypernymy relations in the WordNet taxonomy (Miller, 1995), together with word embeddings (Mikolov et al., 2013; Pennington et al., 2014).

A hypernymy relation represents an asymmetric relation between a class and each of its instances. A hypernym (e.g., *vertebrate*) has a broader context than its hyponyms (*bird*, *fishes*, *reptiles* etc). Conversely, the contextual properties of the hyponyms are usually a subset of those of their hypernym(s). Hypernymy has been used extensively in natural language processing, including in recent works such as Yu et al. (2015) and HyperVec (Nguyen et al., 2017) that have proposed learning word embeddings that reflect the hypernymy relation. Based on this, we have decided to make use of available hypernym-hyponym data to propose an approach for labeling clusters of keywords by a representative selection of their hypernyms.

In the proposed approach, we first extract a set of keywords from the original document collection. We then apply a step of hierarchical clustering on the keywords to partition them into a hierarchy of clusters. To this aim, we represent each keyword as a real-valued vector using pre-trained word embeddings (Pennington et al., 2014) and repeatedly apply a standard clustering algorithm.

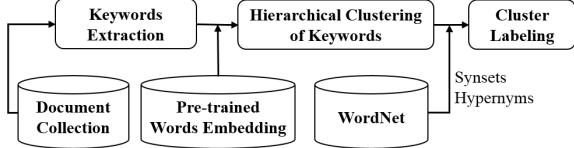


Figure 1: The proposed cluster labeling pipeline.

For labeling the clusters, we first look up all the synonyms of the keywords and, in turn, their hypernyms in the WordNet hierarchy. We then encode the hypernyms as word embeddings and use various approaches to select them based on their distance from the clusters’ centers. The experimental results over a benchmark document collection have shown that such a distance-based selection is reasonably aligned with the hypernyms selected by four, independent human annotators. As a side result, we show that the employed word embeddings spontaneously contain the hypernymy relation, offering a plausible justification for the effectiveness of the proposed method.

## 2 The Proposed Pipeline

The proposed pipeline of processing steps is shown in Figure 1. First, keywords are extracted from each document in turn and accumulated in an overall set of unique keywords. After mapping such keywords to pre-trained word embeddings, hierarchical clustering is applied in a top-down manner. The leaves of the constructed tree are considered as the clusters to be labeled. Finally, each cluster is labeled automatically by leveraging a combination of WordNet’s hypernyms and synsets and word embeddings. The following subsections present each step in greater detail.

### 2.1 Keyword Extraction

For the keyword extraction, we have used the rapid automatic keyword extraction (RAKE) of Rose et al. (2010). This method extracts keywords (i.e., single words or very short word sequences) from a given document collection and its main steps can be summarized as:

1. Split a document into sentences using a pre-defined set of sentence delimiters.
2. Split sentences into sequences of contiguous words at phrase delimiters to build the candidate set.
3. Collect the set of unique words ( $W$ ) that appear in the candidate set.

4. Compute the word co-occurrence matrix  $X_{|W|\times|W|}$  for  $W$ .
5. Calculate word score  $score(w) = deg(w)/freq(w)$ , where  $deg(w) = \sum_{i \in \{1, \dots, |W|\}} X[w, i]$  and  $freq(w) = \sum_{i \in \{1, \dots, |W|\}} (X[w, i] \neq 0)$ .
6. Score each candidate keyword as the sum of its member word scores.
7. Select the top  $T$  scoring candidates as keywords for the document.

Alternatively, RAKE can use other combinations of  $deg(w)$  and  $freq(w)$  as the word scoring function. The keywords extracted from all the documents are accumulated into a set,  $C$ , ensuring uniqueness.

### 2.2 Hierarchical Clustering of Keywords

A top-down approach is used to hierarchically cluster the keywords in  $C$ . First, each component word of each keyword is mapped onto a numerical vector using pre-trained GloVe50d<sup>1</sup> word embeddings (Pennington et al., 2014); missing words are mapped to zero vectors. Then, each keyword  $k$  is represented with the average vector  $\overrightarrow{k}$  of its component words. Then, we start from set  $C$  as the root of the tree and follow a branch-and-bound approach, where each tree node is clustered into  $c$  clusters using the  $k$ -means algorithm (Hartigan and Wong, 1979). A node is marked as a leaf if it contains less than  $n$  keywords or it belongs to level  $d$ , the tree’s depth limit. The leaf nodes are the clusters to be named with a set of verbal terms.

### 2.3 Cluster Labeling

As discussed in Section 1, we aim to label each cluster with descriptive terms. The labels should be more general than the cluster’s members to abstract the nature of the cluster. To this end, we leverage the hypernym-hyponym correspondences in the lexical ontology. First, for each cluster, we create a large set,  $L$ , of candidate labels by including the hypernyms<sup>2</sup> of the component words, expanded by their synonyms, of all the keywords. The synonyms are retrieved from the WordNet’s sets of synonyms, called *synsets*. Then, we apply the four following approaches to select  $l$  labels from set  $L$ :

<sup>1</sup><http://nlp.stanford.edu/data/wordvecs/glove.6B.zip>

<sup>2</sup>Nouns only (not verbs).

- *FreqKey*: Choose the  $l$  most frequent hypernyms of the  $l$  most frequent keywords.
- *CentKey*: Choose the  $l$  most central hypernyms of the  $l$  most central keywords.
- *FreqHyp*: Choose the  $l$  most frequent hypernyms.
- *CentHyp*: Choose the  $l$  most central hypernyms.

Approaches *FreqKey* and *FreqHyp* are based on frequencies in the collection. For performance evaluation, we sort their selected labels in descending frequency order. In *CentKey* and *CentHyp*, the centrality is computed with respect to the cluster’s center in the embedding space as the average vector of all its keywords  $\vec{K} = \frac{1}{|K|} \sum_{k \in K} \vec{k}$ . The distance between hypernym  $h$  and the cluster’s center is  $d(\vec{h}, \vec{K}) = \|\vec{h} - \vec{K}\|$ , where  $\vec{h}$  is the average vector of the hypernym’s component words. The labels selected by these two approaches are sorted in ascending distance order.

### 3 Experiments and Results

For the experiments, we have used the WebAP dataset<sup>3</sup> (Keikha et al., 2014) as the document collection. This dataset contains 6,399 documents of diverse nature with a total of 1,959,777 sentences. For the RAKE software<sup>4</sup>, the hyper-parameters are the minimum number of characters of each keyword, the maximum number of words of each keyword, and the minimum number of times each keyword appears in the text, and they have been left to their default values of 5, 3, and 4, respectively. Likewise, parameter  $T$  has been set to its default value of one third of the words in the co-occurrence matrix. For the hierarchical clustering, we have used  $c = 8$ ,  $n = 100$  and  $d = 4$  based on our own subjective assessment.

#### 3.1 Human Annotation and Evaluation

For the evaluation, eight clusters (one from each sub-tree) were chosen to be labeled manually by four, independent human annotators. For this purpose, for each cluster, we provided the list of its keywords,  $K$ , and the candidate labels,  $L$ , to the annotators, and asked them to select the best  $l = 10$  terms from  $L$  to describe the cluster. Initially,

<sup>3</sup><https://ciir.cs.umass.edu/downloads/WebAP/>

<sup>4</sup><https://github.com/aneesha/RAKE>

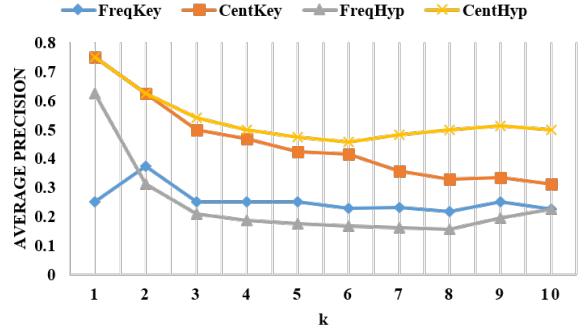


Figure 2: Precision at  $k$  ( $P@k$ ) for  $k = 1, \dots, 10$  averaged over the eight chosen clusters for the compared approaches.

we had considered asking the annotators to also select representative labels from  $K$ , but a preliminary analysis showed that they were unsuitable to describe the cluster as a whole (Table 1 shows an example). Although the annotators were asked to provide their selection as a ranked list, we did not make use of their ranking order in the evaluation.

To evaluate the prediction accuracy, for each cluster we have considered the union of the lists provided by the human annotators as the ground truth (since  $|L|$  was typically in the order of 150 – 200, the intersection of the lists was often empty or minimal). As performance figure, we have decided to report the well-known precision at  $k$  ( $P@k$ ) for values of  $k$  between one and ten. We have not used the recall since the ground truth had size 40 in most cases while the prediction’s size was kept to  $l = 10$  in all cases, resulting in a highest possible recall of 0.25. Figure 2 compares the average  $P@k$  for  $k = 1, \dots, 10$  for the four proposed approaches. The two approaches based on minimum distance to the cluster center (*CentKey* and *CentHyp*) have outperformed the other two approaches based on frequencies (*FreqKey* and *FreqHyp*) for all values of  $k$ . This shows that the word embedding space is in good correspondence with the human judgement. Moreover, approach *CentHyp* has outperformed all other approaches for all values of  $k$ , showing that the hypernyms’ centrality in the cluster is the key property for their effective selection.

#### 3.2 Visualization of Keywords and Hypernyms

Hypernyms are more general terms than the corresponding keywords, thus we expect them to be in larger mutual distance in the word embedding

Keywords	website www, clearinghouse, nih website, bulletin, websites, hotline, kbr publications, pfm file, syst publication, gov web site, dhhs publication, beta site, lexis nexis document, private http, national register bulletin, daily routines, data custodian, information, serc newsletter, certified mail, informational guide, dot complaint database, coverage edit followup, local update, mass mailing, ahrq web site, homepage, journal messenger, npl site, pdf private, htm centers, org website, web site address, telephone directory, service records, page layout program, service invocation, newsletter, card reader, advisory workgroup, library boards, full text online, usg publication, webpage, bulletin boards, fbis online, teleconference info, journal url, insert libraries, headquarters files, volunteer website http, bibliographic records, vch publishers, pid web site, tsbp newsletter, electronic bulletin boards, email addresses, ecommerce, traveler, api service, intranet, website http, newsletter nps files, mail advertisement transmitted, subscribe, mna programs, npci website, bulletin board, fais information, archiving, page attachment, nondriver id, mail etiquette, ip address, national directory, web page, pdq editorial boards, aml sites, dhs site, pid website, directory ers web site, forums, digest, beta site management, directories, ccr papers, iee press, fips publication, org web site, clearinghouse database, monterey database, hotlines, dslip description info, danish desk files, sos web site, bna program, newsletters, inspections portal page, letterhead, app propri, image file directory, website, electronic mail notes, web site http, customized template page, mail addresses, health http, internet questionnaire assistance, electronic bulletin board, eos directly addresses, templates directory, beta site testers, informational, dataplot auxiliary directory, coverage edit, quarterly newsletter, distributed, reader, records service, web pages.
Annotator 1	<b>electronic.communication</b> , computer.network, <b>web.page</b> , <b>web.site</b> , mail, <b>text.file</b> , <b>computer.file</b> , protocol, software, electronic.equipment
Annotator 2	computer.network, telecommunication, computer, mail, <b>web.page</b> , information, news, press, code, software
Annotator 3	news, <b>informing</b> , medium, <b>web.page</b> , <b>computer.file</b> , written.record, document, press, article, essay
Annotator 4	communication, <b>electronic.communication</b> , <b>informing</b> , press, medium, document, electronic.equipment, computer.network, transmission, record
<i>CentHyp</i>	<b>electronic.communication</b> , information.measure, <b>text.file</b> , <b>web.page</b> , <b>informing</b> , print.media, <b>web.site</b> , <b>computer.file</b> , commercial.enterprise, reference.book

Table 1: An example cluster. The hypernyms selected by *CentHyp* and by at least one annotator are shown in boldface.

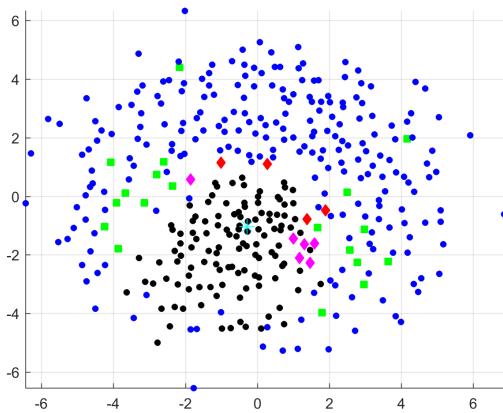


Figure 3: Two-dimensional visualization of an example cluster (this figure should be viewed in color). The black and blue dots are the cluster’s keywords and the keywords’ hypernyms, respectively. The green dots are the hypernyms selected by the human annotators, the red dots are the hypernyms selected by *CentHyp*, and their intersection is recolored in magenta. The cluster’s center is the turquoise star.

space. To explore their distribution, we have used two-dimensional multidimensional scaling (MDS) visualizations (Borg and Groenen, 2005) of selected clusters. For each cluster, the keywords set  $K$ , the hypernyms set  $L$ , and the cluster’s center have all been aggregated as a single set before applying MDS. An example is shown in Figure 3. As can be seen, the hypernyms (blue dots) nicely distribute as a circular crown, external and concentric to the keywords (black dots), showing that the hypernymy relation corresponds empirically to a radial expansion away from the cluster’s center. This likely stems from the embedding space’s requirement to simultaneously enforce meaningful distances between the different keywords, the keywords and the corresponding hypernyms, and between the hypernyms themselves. The hypernyms selected by the annotators (green and magenta

dots) are among the closest to the cluster’s center, and thus those selected by *CentHyp* (red and magenta dots) have the best correspondence (magenta dots alone) among the explored approaches.

### 3.3 A Detailed Example

As a detailed example, Table 1 lists all the keywords of a sample cluster and the hypernyms selected by the four human annotators and *CentHyp*. Some of the hypernyms selected by more than one annotator (e.g., “electronic communication”, “web page” and “computer file”) have also been successfully identified by *CentHyp*. On the other hand, *CentHyp* has selected at least two terms (“commercial enterprise” and “reference book”) that are unrelated to the cluster. Qualitatively, we deem the automated annotation as noticeably inferior to the human annotations, yet usable wherever manual annotation is infeasible or impractical.

## 4 Conclusion

This paper has explored various approaches for labeling keyword clusters based on the hypernyms from the WordNet lexical ontology. The proposed approaches map both the keywords and their hypernyms to a word embedding space and leverage the notion of centrality in the cluster. Experiments carried out using the WebAP dataset have shown that one of the approaches (*CentHyp*) has outperformed all the others in terms of precision at  $k$  for all values of  $k$ , and it has provided labels which are reasonably aligned with those of a pool of annotators. We plan to test the usefulness of the labels for tasks of search expansion in the near future.

## Acknowledgments

This research has been funded by the Capital Markets Cooperative Research Centre in Australia and supported by Semantic Sciences Pty Ltd.

## References

- I. Borg and P. J. F. Groenen. 2005. *Modern Multidimensional Scaling: Theory and Applications*. Springer Series in Statistics. Springer New York.
- J. A. Hartigan and M. A. Wong. 1979. A K-Means Clustering Algorithm. *JSTOR: Applied Statistics* 28(1):100–108.
- M. Keikha, J. H. Park, W. B. Croft, and M. Sanderson. 2014. Retrieving Passages and Finding Answers. In *Proceedings of the 2014 Australasian Document Computing Symposium (ADCS)*, pages 81–84.
- J. H. Lau, K. Grieser, D. Newman, and T. Baldwin. 2011. Automatic Labelling of Topic Models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL)*, volume 1, pages 1536–1545.
- C. D. Manning, P. Raghavan, and H. Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.
- T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems (NIPS)*, volume 2, pages 3111–3119.
- G. A. Miller. 1995. WordNet: A Lexical Database for English. *Communications of the ACM* 38(11):39–41.
- K. A. Nguyen, M. Köeper, S. Schulte im Walde, and N. T. Vu. 2017. Hierarchical Embeddings for Hyponymy Detection and Directionality. In *Proceedings of the 2017 Empirical Methods in Natural Language Processing (EMNLP)*, pages 233–243.
- J. Pennington, R. Socher, and C. D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Empirical Methods in Natural Language Processing (EMNLP)*, volume 14, pages 1532–1543.
- S. Rose, D. Engel, N. Cramer, and W. Cowley. 2010. Automatic Keyword Extraction from Individual Documents. In *Text Mining. Applications and Theory*, Wiley-Blackwell, chapter 1, pages 1–20.
- J. Wang, C. Kang, Y. Chang, and J. Han. 2014. A Hierarchical Dirichlet Model for Taxonomy Expansion for Search Engines. In *Proceedings of the 23rd International Conference on World Wide Web (WWW)*, pages 961–970.
- Z. Yu, H. Wang, X. Lin, and M. Wang. 2015. Learning Term Embeddings for Hyponymy Identification. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1390–1397.

# A Comparative Study of Embedding Models in Predicting the Compositionality of Multiword Expressions

Navnita Nandakumar

Bahar Salehi

Timothy Baldwin

School of Computing and Information Systems

The University of Melbourne

Victoria 3010, Australia

nnandakumar@student.unimelb.edu.au

{salehi.b,tbaldwin}@unimelb.edu.au

## Abstract

In this paper, we perform a comparative evaluation of off-the-shelf embedding models over the task of compositionality prediction of multiword expressions (“MWEs”). Our experimental results suggest that character- and document-level models do capture some aspects of MWE compositionality and are effective at modelling varying levels of compositionality, but ultimately are not as effective as a simple word2vec baseline. However they have the advantage over word-level models that they do not require token-level identification of MWEs in the training corpus.

## 1 Introduction

In recent years, the study of the semantic idiomacity of multiword expressions (“MWEs”: Baldwin and Kim (2010)) has focused on *compositionality prediction*, a regression task involving the mapping of an MWE onto a continuous scale, representing its compositionality either as a whole or for each of its component words (Reddy et al., 2011; Ramisch et al., 2016; Cordeiro et al., to appear). In the case of *couch potato* “an idler who spends much time on a couch (usually watching television)”, e.g., on a scale of [0, 1] the overall compositionality may be judged to be 0.3, and the compositionality of *couch* and *potato* as 0.8 and 0.1, respectively. The main motivation for the study of compositionality is to better understand the semantic of the compound and the semantic relationships between the component words of the MWEs, which has applications in various information retrieval and natural language processing tasks (Venkatapathy and Joshi, 2006; Acosta et al., 2011; Salehi et al., 2015b).

Separately, there has been burgeoning interest

in learning distributed representations of words and their meanings, starting out with word embeddings (Mikolov et al., 2013; Pennington et al., 2014) and now also involving the study of character- and document-level models (Baroni et al., 2014; Le and Mikolov, 2014; Bojanowski et al., 2017; Conneau et al., 2017). This work has been applied in part to predicting the compositionality of MWEs (Salehi et al., 2015a; Hakimi Parizi and Cook, 2018), work that this paper builds on directly, in performing a comparative study of the performance of a range of off-the-shelf representation learning methods over the task of MWE compositionality prediction.

Our contributions are as follows: (1) we show that, despite their effectiveness over a range of other tasks, recent off-the-shelf character- and document-level embedding learning methods are inferior to simple word2vec at modelling MWE compositionality; and (2) we demonstrate the utility of using paraphrase data in addition to simple lemmas in predicting MWE compositionality.

## 2 Related work

The current state-of-the-art in compositionality prediction involves the use of word embeddings (Salehi et al., 2015a). The vector representations of each component word (e.g. *couch* and *potato*) and the overall MWE (e.g. *couch potato*) are taken as a proxy for their respective meanings, and compositionality of the MWE is then assumed to be proportional to the relative similarity between each of the components and overall MWE embedding. However, word-level embeddings require token-level identification of each MWE in the training corpus, meaning that if the set of MWEs changes, the model needs to be retrained. This limitation led to research on character-level models, since character-level models can implic-

itly handle an unbounded vocabulary of component words and MWEs (Hakimi Parizi and Cook, 2018). There has also been work in the extension of word embeddings to document embeddings that map entire sentences or documents to vectors (Le and Mikolov, 2014; Conneau et al., 2017).

### 3 Embedding Methods

We use two character-level embedding models (`fastText` and `ELMo`) and two document-level models (`doc2vec` and `inferSent`) to compare with word-level `word2vec`, as used in the state-of-the-art method of Salehi et al. (2015a). In each case, we use canonical pre-trained models, with the exception of `word2vec`, which must be trained over data with appropriate tokenisation to be able to generate MWE embeddings, as it treats words atomically and cannot generate OOV words.

#### 3.1 Word-level Embeddings

Word embeddings are mappings of words to vectors of real numbers. This helps create a more compact (by means of dimensionality reduction) and expressive (by means of contextual similarity) word representation.

**word2vec** We trained `word2vec` (Mikolov et al., 2013) over the latest English Wikipedia dump.<sup>1</sup> We first pre-processed the corpus, removing XML formatting, stop words and punctuation, to generate clean, plain text. We then iterated through 1% of the corpus (following Hakimi Parizi and Cook (2018)) to find every occurrence of each MWE in our datasets and concatenate them, assuming every occurrence of the component words in sequence to be the compound noun (e.g. every *couch potato* in the corpus becomes *couchpotato*). We do this because instead of a single embedding for the MWE, `word2vec` generates separate embeddings for each of the component words, owing to the space between them. If the model still fails to generate embeddings for either the MWE or its components (due to data sparseness), we assign the MWE a default compositionality score of 0.5 (neutral). In the case of paraphrases, we compute the element-wise average of the embeddings of each of the component words to generate the embedding of the phrase.

---

<sup>1</sup>Dated 02-Oct-2018, 07:23

#### 3.2 Character-level Embeddings

In a character embedding model, the vector for a word is constructed from the character  $n$ -grams that compose it. Since character  $n$ -grams are shared across words, assuming a closed-world alphabet,<sup>2</sup> these models can generate embeddings for OOV words, as well as words that occur infrequently. The two character-level embedding models we experiment with are `fastText` (Bojanowski et al., 2017) and `ELMo` (Peters et al., 2018), as detailed below.

**fastText** We used the 300-dimensional model pre-trained on Common Crawl and Wikipedia using CBOW. `fastText` assumes that all words are whitespace delimited, so in order to generate a representation for the combined MWE, we remove any spaces and treat it as a fused compound (e.g. *couch potato* becomes *couchpotato*). In the case of paraphrases, we use the same word averaging technique as we did in `word2vec`.

**ELMo** We used the ElmoEmbedder class in Python’s `allennlp` library.<sup>3</sup> The model was pre-trained over SNLI and SQuAD, with a dimensionality of 1024.

Note that the primary use case of `ELMo` is to generate embeddings in context, but we are not providing any context in the input, for consistency with the other models. As such, we are knowingly not harnessing the full potential of the model. However, this naive use of `ELMo` is not inappropriate as the relative compositionality of a compound is often predictable from its component words only, even for novel compounds such as *giraffe potato* (which has a plausible compositional interpretation, as a potato shaped like a giraffe) vs. *couch intelligence* (where there is no natural interpretation, suggesting that it may be non-compositional).

#### 3.3 Document-level Embeddings

Document-level embeddings aim to learn vector representations of documents (sentences or even paragraphs), to generate a representation

---

<sup>2</sup>Which is a safe assumption for languages with small-scale alphabetic writing systems such as English, but potentially problematic for languages with large orthographies such as Chinese (with over 10k ideograms in common use, and many more rarer characters) or Korean (assuming we treat each Hangul syllable as atomic).

<sup>3</sup>`options_file = https://bit.ly/2CInZPV,`  
`weight_file = https://bit.ly/2PvNqHh`

of its overall content in the form of a fixed-dimensionality vector. The two document-level embeddings used in this research are `doc2vec` (Le and Mikolov, 2014) and `infersent` (Conneau et al., 2017), as detailed below.

**doc2vec** We used the gensim implementation of `doc2vec` (Lau and Baldwin, 2016; Řehůřek and Sojka, 2010), pretrained on Wikipedia data using the `word2vec` skip-gram models pretrained on Wikipedia and AP News.<sup>4</sup>

**infersent** We used two versions of `infersent` of 300 dimensions, using the inbuilt `infersent.build_vocab_k_words` function to train the model over the 100,000 most popular English words, using: (1) `GloVe` (Pennington et al., 2014) word embeddings (“`infersentGloVe`”); and (2) `fastText` word embeddings (“`infersentfastText`”).

## 4 Modelling Compositionality

In order to measure the overall compositionality of an MWE, we propose the following three broad approaches.

### 4.1 Direct Composition

Our first approach is to directly compare the embeddings of each of the component nouns with the embedding of the MWE via cosine similarity, in one of two ways: (1) pre-combine the embeddings for the component words via element-wise sum, and compare with the embedding for the MWE (“ $\text{Direct}_{\text{pre}}$ ”); and (2) compare each individual component word with the embedding for the MWE, and post-hoc combine the scores via a weighted sum (“ $\text{Direct}_{\text{post}}$ ”). Formally:

$$\begin{aligned} \text{Direct}_{\text{pre}} &= \cos(\mathbf{mwe}, \mathbf{mwe}_1 + \mathbf{mwe}_2) \\ \text{Direct}_{\text{post}} &= \alpha \cos(\mathbf{mwe}, \mathbf{mwe}_1) + \\ &\quad (1 - \alpha) \cos(\mathbf{mwe}, \mathbf{mwe}_2) \end{aligned}$$

where:  $\mathbf{mwe}$ ,  $\mathbf{mwe}_1$ , and  $\mathbf{mwe}_2$  are the embeddings for the combined MWE, first component and second component, respectively;<sup>5</sup>  $\mathbf{mwe}_1 + \mathbf{mwe}_2$  is the element-wise sum of the vectors of each of the component words of the MWE; and  $\alpha \in [0, 1]$  is a scalar which allows us to vary the weight of

<sup>4</sup><https://github.com/jhlau/doc2vec/blob/master/README.md>

<sup>5</sup>Noting that all MWEs are binary in our experiments, but equally that the methods generalise trivially to larger MWEs.

Emb. method	$\text{Direct}_{\text{pre}}$	$\text{Direct}_{\text{post}}$
word2vec	<b>0.684</b>	<b>0.710</b> ( $\alpha = 0.3$ )
fastText	0.223	0.285 ( $\alpha = 0.3$ )
ELMo	0.056	0.399 ( $\alpha = 0.0$ )
doc2vec	-0.049	0.025 ( $\alpha = 0.0$ )
<code>infersent<sub>GloVe</sub></code>	0.413	0.500 ( $\alpha = 0.5$ )
<code>infersent<sub>infersent</sub></code>	0.557	0.610 ( $\alpha = 0.5$ )

Table 1: Pearson correlation coefficient for compositionality prediction results on the REDDY dataset.

the respective components in predicting the compositionality of the compound. The intuition behind both of these methods is that if the MWE appears in similar contexts to its components, then it is compositional.

### 4.2 Paraphrases

Our second approach is to calculate the similarity of the MWE embedding with that of its paraphrases, assuming that we have access to paraphrase data.<sup>6</sup> We achieve this using the following three formulae:

$$\text{Para\_first} = \cos(\mathbf{mwe}, \mathbf{para}_1)$$

$$\text{Para\_all}_{\text{pre}} = \cos(\mathbf{mwe}, \sum_i \mathbf{para}_i)$$

$$\text{Para\_all}_{\text{post}} = \frac{1}{N} \sum_{i=1}^N \cos(\mathbf{mwe}, \mathbf{para}_i)$$

where  $\mathbf{para}_1$  and  $\mathbf{para}_i$  denote the embedding for the first (most popular) and  $i$ -th paraphrases, respectively.

We apply this method to RAMISCH only, since REDDY does not have any paraphrase data (see Section 5.1 for details).

### 4.3 Combination

Our final approach (“Combined”) is based on the combination of the direct composition and paraphrase methods, as follows:

$$\begin{aligned} \text{Combined} &= \beta \max (\text{Direct}_{\text{pre}}, \text{Direct}_{\text{post}}) + \\ &\quad (1 - \beta) \max (\text{Para\_first}, \text{Para\_all}_{\text{pre}}, \\ &\quad \quad \quad \text{Para\_all}_{\text{post}}) \end{aligned}$$

where  $\beta \in [0, 1]$  is a scalar weighting factor to balance the effects of the two methods. The choice

<sup>6</sup>Each paraphrase shows an interpretation of the compound semantics. e.g. *olive oil* is “oil from olive”

Emb. method	Direct <sub>pre</sub>	Direct <sub>post</sub>	Para_first	Para_all <sub>pre</sub>	Para_all <sub>post</sub>	Combined
word2vec	<b>0.667</b>	<b>0.731</b> ( $\alpha = 0.7$ )	<b>0.714</b>	<b>0.822</b>	<b>0.880</b>	<b>0.880</b> ( $\beta = 0.0$ )
fastText	0.395	0.446 ( $\alpha = 0.7$ )	0.569	0.662	0.704	0.704 ( $\beta = 0.0$ )
ELMo	0.139	0.295 ( $\alpha = 0.0$ )	0.367	0.642	0.664	0.669 ( $\beta = 0.2$ )
doc2vec	-0.146	0.048 ( $\alpha = 1.0$ )	0.405	0.372	0.401	0.419 ( $\beta = 0.3$ )
inferent <sub>GloVe</sub>	0.321	0.427 ( $\alpha = 0.7$ )	0.639	0.704	0.741	0.774 ( $\beta = 0.5$ )
inferent <sub>fastText</sub>	0.274	0.380 ( $\alpha = 0.8$ )	0.615	0.781	0.783	0.783 ( $\beta = 0.0$ )

Table 2: Pearson correlation coefficient for compositionality prediction results on the RAMISCH dataset.

of the max operator here to combine the sub-methods for each of the direct composition and paraphrase methods is that all methods tend to underestimate the compositionality (and empirically, it was superior to taking the mean).

## 5 Experiments

### 5.1 Datasets

We evaluate the models on the following two datasets, which are comprised of 90 English binary noun compounds each, rated for compositionality on a scale of 0 (non-compositional) to 5 (compositional). In each case, we evaluate model performance via the Pearson’s correlation coefficient ( $r$ ).

**REDDY** This dataset contains scores for the compositionality of the overall MWE, as well as that of each component word (Reddy et al., 2011); in this research, we use the overall compositionality score of the MWE only, and ignore the component scores.

**RAMISCH** Similarly to REDDY, this dataset contains scores for the overall compositionality of the MWE as well as the relative compositionality of each of its component words, in addition to paraphrases suggested by the annotators, in decreasing order of popularity (Ramisch et al., 2016); in this research, we use the overall compositionality score and paraphrase data only.

### 5.2 Results and Discussion

The results of the experiments on REDDY and RAMISCH are presented in Tables 1 and 2, respectively. In this work, we simplistically present the results for the best  $\alpha$  and  $\beta$  values for each method over a given dataset, meaning we are effectively peaking at our test data. Sensitivity of the  $\alpha$  hyper-parameter is shown in Figures 1 and

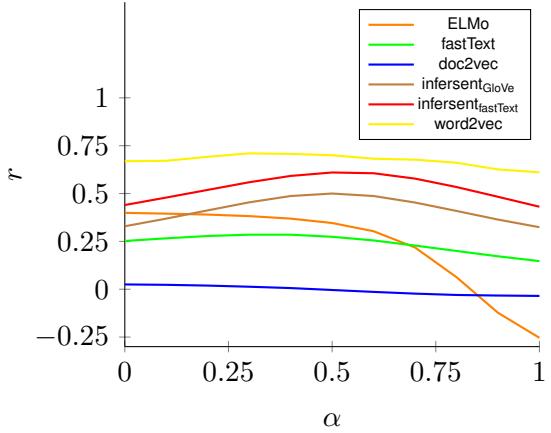


Figure 1: Sensitivity analysis of  $\alpha$  (REDDY)

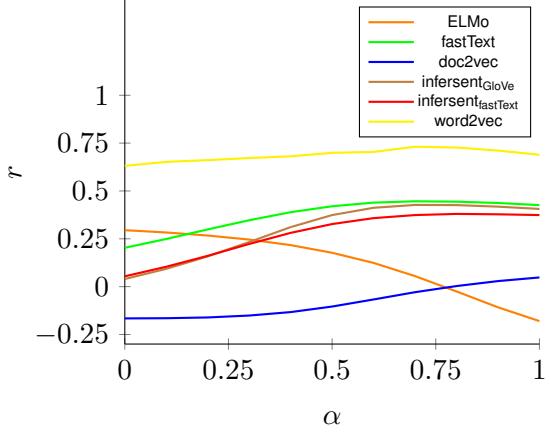


Figure 2: Sensitivity analysis of  $\alpha$  (RAMISCH)

2, for the REDDY and RAMISCH datasets, respectively.

The first observation to be made is that none of the pretrained models match the state-of-the-art method based on word2vec, despite the simplicity of the method. ELMo and doc2vec in particular perform worse than expected, suggesting that their ability to model non-compositional language is limited. Recall, however, our comment about

using ELMo naively, in not including any context when generating the embeddings for the component words and, more importantly, the overall MWE. The results show that doc2vec performs better when representing paraphrases, and struggles with compounds without sentential context.

In Table 1, we find Direct<sub>post</sub> to produce a higher correlation in all cases, with  $\alpha$  ranging from 0.0 to 0.5, suggesting that the second element (= head) contributes more to the overall compositionality of the MWE than the first element (= modifier); this is borne out in Figure 1.

In Table 2, on the other hand, we find that, with the exception of ELMo, the  $\alpha$  values favour the modifier of the MWE over the head (i.e.  $\alpha > 0.5$ ; also seen in Figure 2), implying that the former is more significant in predicting the compositionality of the MWE. The reason for the mismatch between the two datasets is not immediately clear, other than the obvious data sparsity.

We also see that the paraphrases achieve a higher correlation across all models, suggesting this is a promising direction for future study. The low  $\beta$  values for Combined also confirm that the paraphrase methods have greater predictive power than the direct composition methods. Among the paraphrase experiments, we find that Para\_all<sub>post</sub>—the average of the similarities of the MWE with each of its paraphrases—consistently achieves the best results. We hypothesize that the paraphrases provide additional information regarding the compounds that further help determine their compositionality.

## 6 Conclusions and Future Work

This paper has investigated the application of a range of embedding generation methods to the task of predicting the compositionality of an MWE, either directly based on the MWE and its component words, or indirectly based on paraphrase data for the MWE. Our results show that modern character- and document-level embedding models are inferior to the simple word2vec approach at the task. We also show that paraphrase data captures valuable data regarding the compositionality of the MWE.

Since we have achieved such promising results with the paraphrase data, it might be interesting to consider other possible settings in future tests. While none of the other approaches could outperform word2vec, it is useful to note that they were

pretrained and, as such, did not require any manipulation of the training corpus in order to generate vector embeddings of the MWEs. This means they can be applied to new datasets without the need for retraining and are, therefore, more robust.

In future work, we intend to train the models used in our study on a fixed corpus, to compare their performance in a more controlled setting. We will also do proper tuning of the hyperparameters over held-out data, and plan to experiment with other languages.

## References

- Otavio Acosta, Aline Villavicencio, and Viviane Moreira. 2011. Identification and treatment of multiword expressions applied to information retrieval. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*. Portland, USA, pages 101–109.
- Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. In Nitin Indurkha and Fred J. Damerau, editors, *Handbook of Natural Language Processing*, CRC Press, Boca Raton, USA. 2nd edition.
- Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don’t count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014)*. Baltimore, USA, pages 238–247.
- Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics* 5:135–146.
- Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. pages 670–680.
- Silvio Cordeiro, Aline Villavicencio, Marco Idiart, and Carlos Ramisch. to appear. Unsupervised compositionality prediction of nominal compounds. *Computational Linguistics*.
- Ali Hakimi Parizi and Paul Cook. 2018. Do character-level neural network language models capture knowledge of multiword expression compositionality? In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*. pages 185–192.
- Jey Han Lau and Timothy Baldwin. 2016. An empirical evaluation of doc2vec with practical insights into document embedding generation. In *Proceedings of the 1st Workshop on Representation Learning for NLP*. pages 78–86.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning (ICML 2014)*. Beijing, China, pages 1188–1196.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of Workshop at the International Conference on Learning Representations*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*. pages 1532–1543.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. pages 2227–2237.

Carlos Ramisch, Silvio Cordeiro, Leonardo Zilio, Marco Idiart, and Aline Villavicencio. 2016. How naked is the naked truth? a multilingual lexicon of nominal compound compositionality. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. pages 156–161.

Siva Reddy, Diana McCarthy, and Suresh Manandhar. 2011. An empirical study on compositionality in compound nouns. In *Proceedings of 5th International Joint Conference on Natural Language Processing*. pages 210–218.

Radim Řehůřek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. pages 45–50.

Bahar Salehi, Paul Cook, and Timothy Baldwin. 2015a. A word embedding approach to predicting the compositionality of multiword expressions. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. pages 977–983.

Bahar Salehi, Nitika Mathur, Paul Cook, and Timothy Baldwin. 2015b. The impact of multiword expression compositionality on machine translation evaluation. In *Proceedings of the NAACL HLT 2015 Workshop on Multiword Expressions*. Denver, USA, pages 54–59.

Sriram Venkatapathy and Aravind K Joshi. 2006. Using information about multi-word expressions for the word-alignment task. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*. pages 20–27.

# Towards Efficient Machine Translation Evaluation by Modelling Annotators

Nitika Mathur      Timothy Baldwin      Trevor Cohn

School of Computing and Information Systems

The University of Melbourne

Victoria 3010, Australia

nmathur@student.unimelb.edu.au      {tbaldwin,tcohn}@unimelb.edu.au

## Abstract

Current machine translation evaluations use Direct Assessment, based on crowd-sourced judgements from a large pool of workers, along with quality control checks, and a robust method for combining redundant judgements. In this paper we show that the quality control mechanism is overly conservative, increasing the time and expense of the evaluation. We propose a model that does not filter workers, and takes into account varying annotator reliabilities. Our model effectively weights each worker’s scores based on the inferred precision of the worker, and is much more reliable than the mean of either the raw or standardised scores.

## 1 Introduction

Accurate evaluation is critical for measuring progress in machine translation (MT). Despite progress over the years, automatic metrics are still biased, and human evaluation is still a fundamental requirement for reliable evaluation. The process of collecting human annotations is time-consuming and expensive, and the data is always noisy. The question of how to efficiently collect this data has evolved over the years, but there is still scope for improvement. Furthermore, once the data has been collected, there is no consensus on the best way to reason about translation quality.

Direct Assessment (“DA”: Graham et al. (2017)) is currently accepted as the best practice for human evaluation, and is the official method at the Conference for Machine Translation (Bojar et al., 2017a). Every annotator scores a set of translation-pairs, which includes quality control items designed to filter out unreliable workers.

However, the quality control process has low recall for good workers: as demonstrated in Section 3, about one third of good data is discarded, increasing expense. Once good workers are identified, their outputs are simply averaged to produce the final ‘true’ score, despite their varying accuracy.

In this paper, we provide a detailed analysis of these shortcomings of DA and propose a Bayesian model to address these issues. Instead of standardising individual worker scores, our model can automatically infer worker offsets using the raw scores of all workers as input. In addition, by learning a worker-specific precision, each worker effectively has a differing magnitude of vote in the ensemble. When evaluated on the WMT 2016 Tr-En dataset which has a high proportion of unskilled annotators, these models are more efficient than the mean of the standardised scores.

## 2 Background

The Conference on Machine Translation (WMT) annually collects human judgements to evaluate the MT systems and metrics submitted to the shared tasks. The evaluation methodology has evolved over the years, from 5 point adequacy and fluency rating, to relative rankings (“RR”), to DA. With RR, annotators are asked to rank translations of 5 different MT systems. In earlier years, the final score of a system was the expected number of times its translations score better than translations by other systems (expected wins). Bayesian models like Hopkins and May (Hopkins and May, 2013) and Trueskill (Sakaguchi et al., 2014) were then proposed to learn the relative ability of the MT systems. Trueskill was adopted by WMT in 2015 as it is more stable and efficient than the expected wins heuristic.

DA was trialled at WMT 2016 (Bojar et al.,

2016a), and has replaced RR since 2017 (Bojar et al., 2017a). It is more scalable than RR as the number of systems increases (we need to obtain one annotation per system, instead of one annotation per system pair). Each translation is rated independently, minimising the risk of being influenced by the relative quality of other translations. Ideally, it is possible that evaluations can be compared across multiple datasets. For example, we can track the progress of MT systems for a given language pair over the years.

Another probabilistic model, EASL (Sakaguchi and Van Durme, 2018), has been proposed that combines some advantages of DA with Trueskill. Annotators score translations from 5 systems at the same time on a sliding scale, allowing users to explicitly specify the magnitude of difference between system translations. Active learning to select the systems in each comparison to increase efficiency. But it does not model worker reliability, and is, very likely, not compatible with longitudinal evaluation, as the systems are effectively scored relative to each other.

In NLP, most other research on learning annotator bias and reliability has been on categorical data (Snow et al., 2008; Carpenter, 2008; Hovy et al., 2013; Passonneau and Carpenter, 2014).

### 3 Direct Assessment

To measure adequacy, in DA, annotators are asked to rate how adequately an MT output expresses the meaning of a reference translation using a continuous slider, which maps to an underlying scale of 0–100. These annotations are crowdsourced using Amazon Mechanical Turk, where “workers” complete “Human Intelligence Tasks” (HITs) in the form of one or more micro-tasks.

Each HIT consists of 70 MT system translations, along with an additional 30 control items:

1. degraded versions of 10 of these translations;
2. 10 reference translations by a human expert, corresponding to 10 system translations; and
3. repeats of another 10 translations.

The scores on the quality control items are used to filter out workers who either click randomly or on the same score continuously. A conscientious worker would give a near perfect score to reference translations, give a lower score to degraded translations when compared to the corresponding MT system translation, and be consistent with scores for repeat translations.

The paired Wilcoxon rank-sum test is used to test whether the worker scored degraded translations worse than the corresponding system translation. The (arbitrary but customary) cutoff of  $p < 0.05$  is used to determine good workers. The paired Wilcoxon rank-sum test ( $p < 0.05$ ) is used to test whether the worker scored degraded translations worse than the corresponding system translation. The remaining workers are further tested to check that there is no significant difference between their scores for repeat-pairs.

Worker scores are manually examined to filter out workers who obviously gave the same score to all translations, or scored translations at random. Only these workers are rejected payment. Thus, other workers who do not pass the quality control check are paid for their efforts, but their scores are unused, increasing the overall cost.

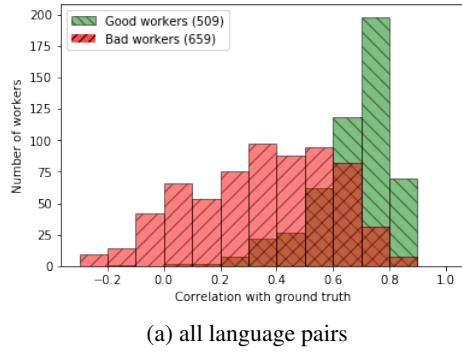
Some workers might have high standards and give consistently low scores for all translations, while others are more lenient. And some workers may only use the central part of the scale. Standardising individual workers’ scores makes them more comparable, and reduces noise before calculating the mean.

The final score of an MT system is the mean standardised score of its translations after discarding scores that do not meet quality control criteria. The noise in worker scores is cancelled out when a large number of translations are averaged.

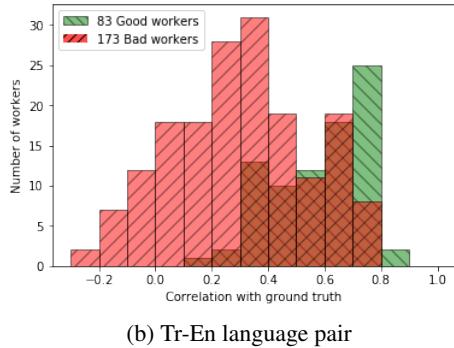
To obtain accurate scores of individual translations, multiple judgments are collected and averaged. As we increase the number of annotators per translation, there is greater consistency and reliability in the mean score. This was empirically tested by showing that there is high correlation between the mean of two independent sets of judgments, when the sample size is greater than 15 (Graham et al., 2015).

However, both these tests are based on a sample-size of 10 items, and, as such, the first test has low power; we show that it filters out a large proportion of the total workers. One solution would be to increase the sample size of the degraded-reference-pairs, but this would be at the expense of the number of useful worker annotations. It is better to come up with a model that would use the scores of all workers, and is more robust to low quality scores.

Automatic metrics such as BLEU (Papineni et al., 2002) are generally evaluated using the Pear-



(a) all language pairs



(b) Tr-En language pair

Figure 1: Accuracy of “good” vs “bad” workers in the WMT 2016 dataset.

son correlation with the mean standardised score of the good workers. We similarly evaluate a worker’s accuracy using the Pearson correlation of the worker’s scores with this ground truth. Over all the data collected for WMT16, the group of good workers are, on average, more accurate than the group of workers who failed the significance test. However, as seen in Figure 1a, there is substantial overlap in the accuracies of the two groups. We can see that very few inaccurate workers were included. However, about a third of the total workers whose scores have a correlation greater than 0.6 were not approved. In particular, over the Tr-En Dataset, the significance test was not very effective, as seen in Figure 1b.

Workers whose scores pass the quality control check are given equal weight, despite the variation in their reliability. Given that quality control is not always reliable (as with the Tr-En dataset, e.g.), this could include worker with scores as low as  $r = 0.2$  correlation with the ground truth.

While worker standardisation succeeds in increasing inter-annotator consistency, this process discards information about the absolute quality of the translations in the evaluation set. When using the mean of standardised scores, we cannot compare MT systems across independent evalua-

tions. In the evaluation of the WMT 17 Neural MT Training Task, the baseline system trained on 4GB GPU memory was evaluated separately from the baseline trained on 8 GB GPU memory and the other submissions. In this setup of manual evaluation, Baseline-4GB scores slightly higher than Baseline-8GB when using raw scores, which is possibly due to chance. However, it scores significantly higher when using standardised scores, which goes against our expectations (Bojar et al., 2017b).

## 4 Models

We use a simple model, assuming that a worker score is normally distributed around the true quality of the translation. Each worker has a precision parameter  $\tau$  that models their accuracy: workers with high  $\tau$  are more accurate. In addition, we include a worker-specific offset  $\beta$ , which models their deviation from the true score.

For each translation  $i \in T$ , we draw the true quality  $\mu$  from the standard normal distribution.<sup>1</sup> Then for each worker  $j \in W$ , we draw their accuracy  $\tau_j$  from a gamma distribution with shape parameter  $k$  and rate parameter  $\theta$ .<sup>2</sup> The offset  $\beta_j$  is again drawn from the standard normal distribution. The worker’s score  $r_{ij}$  is drawn from a normal distribution, with mean  $\mu_i + \beta_j$ , and precision  $\tau_j$ .

$$r_{ij} = \mathcal{N}(\mu_i + \beta_j, \tau_j^{-1}) \quad (1)$$

To help the model, we add constraints on the quality control items: the true quality of the degraded translation is lower than the quality of the corresponding system translation. In addition, the true quality of the repeat items should be approximately equal.

We expect that the model will learn a high  $\tau$  for good quality workers, and give their scores higher weight when estimating the mean. We believe that the additional constraints will help the model to infer the worker precision.

DA can be viewed as the Maximum Likelihood Estimate of this model, with the following substitutions in Equation (1):  $s_{ij}$  is the standardised score of worker  $j$ ,  $\beta_j$  is 0 for all workers, and  $\tau$  is

<sup>1</sup>We first standardise scores (across all workers together) in the dataset

<sup>2</sup>We use  $k = 2$  and  $\theta = 1$  based on manual inspection of the distribution of worker precisions on a development dataset (WMT18 Cs-En)

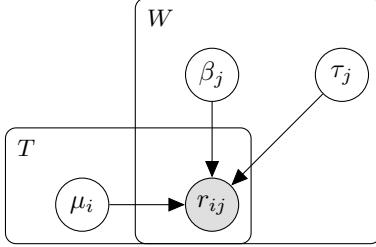


Figure 2: The proposed model, where worker  $j \in W$  has offset  $\beta_j$  and precision  $\tau_j$ , translation  $i \in T$  has quality  $\mu_i$ , and worker  $j$  scores translation  $i$  with  $r_{ij}$

constant for all workers.

$$s_{ij} = \mathcal{N}(\mu_i, \tau^{-1}) \quad (2)$$

The choice of a Gaussian distribution to model worker scores is technically deficient as a Gaussian is unbounded, but it is still a reasonable approximation. This could be remedied, for example, by using a truncated Gaussian distribution, which we leave to future work.

We want to maximise the likelihood of the observed judgments:

$$\begin{aligned} P(r) &= \prod_{j=1}^W \int P(\beta_j) P(\tau_j) \int_{i=1}^T P(\mu_i) \\ &\quad P(r_{i,j} | \mu_i, \beta, \tau) d\beta d\tau d\mu \\ &= \prod_{j=1}^W \mathcal{N}(\beta_j | 0, 1) \Gamma(\tau_j | k, \theta) \int_{i=1}^T \mathcal{N}(\mu_i | 0, 1) \\ &\quad \mathcal{N}(r_{i,j} | \mu_i, \tau^{-1}) d\beta d\tau d\mu \quad (3) \end{aligned}$$

We use the Expectation Propagation algorithm (Minka, 2001) to infer posteriors over  $\mu$  and worker parameters  $\beta$  and  $\tau$ .<sup>3</sup> Expectation Propagation is a technique for approximating distributions which can be written as a product of factors. It iteratively refines each factor by minimising the KL divergence from the approximate to the true distribution.

## 5 Experiments

We evaluate our models on data from the segment-level WMT 16 dataset (Bojar et al., 2016b). We choose the Turkish to English (Tr-En) dataset, which consists of 256 workers, of which about

<sup>3</sup>We use the Infer.NET (Minka et al., 2018) framework to implement our models.

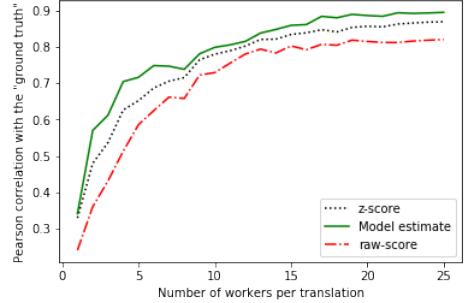


Figure 3: Pearson’s  $r$  of the estimated true score with the “ground truth” as we increase the number of workers per translation.

two thirds (67.58%) fail the quality control measures. It consists of 560 translations, with at least 15 “good” annotations for each of these translations (see Figure 1b).

We use the mean of 15 good standardised annotations as a proxy for the gold standard when evaluating efficiency, and starting from one worker, increase the number of workers to the maximum available. Figure 3 shows that our models are consistently more accurate than the mean of the standardised scores.

Figure 4 shows the learned precision and offset for 5 annotators per translation, against the precision and offset of worker scores calculated with respect to the “ground truth”. This shows that the model is learning worker parameters even when the number of workers is very small, and is using this information to get a better estimate of the mean (the model obtains  $r = 0.72$ , compared to  $r = 0.65$  for the mean  $z$ -score).

On further examination of the outlier in Figure 4a, we find that this worker is pathologically bad. They give a 0 score for all the translations in one HIT, and mostly 100s to the other half. This behaviour is not captured by our model.

## 6 Discussion and Future Work

We showed that significance tests over a small set of quality control items are ineffective at identifying good and bad workers, and propose a model that does not depend on this step. Instead, it uses constraints on the quality control items to learn worker precision, and returns a more reliable estimate of the mean using fewer worker scores per translation. This model does not tell us when to stop collecting judgments. It would be useful to know to have a method to determine when to stop

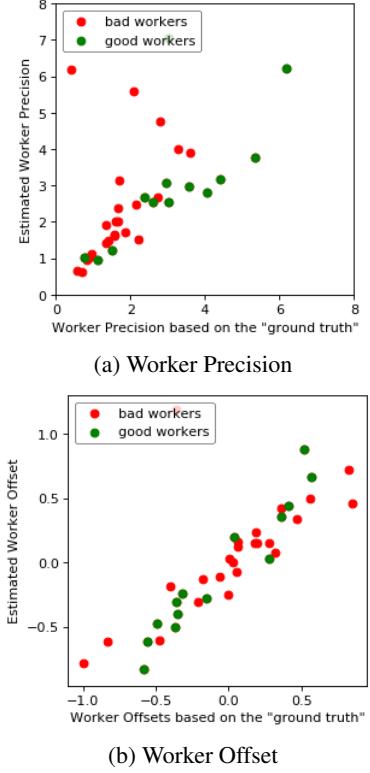


Figure 4: Scatter plot of worker precision/offset inferred by the model with only 5 workers per translation, against the precision/offset of the deltas of the worker score and the “ground truth”.

collecting annotations based on scores received, instead of relying on a number obtained from one-time experiments.

More importantly, we need to have ways to calibrate worker scores to ensure consistent evaluations across years, so we can measure progress in MT over time. Even if a better model is found to calibrate workers, this does not ensure consistency in judgments, and we believe the HIT structure needs to be changed. We propose to replace the 30 quality control items with items of reliably known quality from the previous year. The correlation between the worker scores and the known scores can be used to assess the reliability of the worker. Moreover, we can scale the worker scores based on these known items, to ensure consistent scores over years.

## Acknowledgements

We thank the anonymous reviewers for their valuable feedback and suggestions. This work was supported in part by the Australian Research Council.

## References

- Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017a. Findings of the 2017 Conference on Machine Translation (WMT17). In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*. Copenhagen, Denmark, pages 169–214. <http://www.aclweb.org/anthology/W17-4717>.
- Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016a. Findings of the 2016 Conference on Machine Translation. In *Proceedings of the First Conference on Machine Translation*. Berlin, Germany, pages 131–198.
- Ondřej Bojar, Yvette Graham, Amir Kamran, and Miloš Stanojević. 2016b. Results of the WMT16 Metrics Shared Task. In *Proceedings of the First Conference on Machine Translation*. Berlin, Germany, pages 199–231.
- Ondřej Bojar, Jindřich Helcl, Tom Kocmi, Jindřich Libovický, and Tomáš Musil. 2017b. Results of the WMT17 Neural MT Training task. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*. Copenhagen, Denmark, pages 525–533. <http://www.aclweb.org/anthology/W17-4757>.
- Bob Carpenter. 2008. Multilevel Bayesian models of categorical data annotation. Technical report, Alias-i.
- Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2017. Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering* 23(1):330.
- Yvette Graham, Nitika Mathur, and Timothy Baldwin. 2015. Accurate evaluation of segment-level machine translation metrics. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics Human Language Technologies*. Denver, USA, pages 1183–1191.
- Mark Hopkins and Jonathan May. 2013. Models of translation competitions. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*. Sofia, Bulgaria, pages 1416–1424.
- Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard H Hovy. 2013. Learning whom to trust

with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2013)*. Atlanta, USA, pages 1120–1130.

T. Minka, J.M. Winn, J.P. Guiver, Y. Zaykov, D. Fabian, and J. Bronskill. 2018. Infer.NET 0.3. Microsoft Research Cambridge. <http://dotnet.github.io/infer>.

Thomas P. Minka. 2001. Expectation propagation for approximate Bayesian inference. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*. Seattle, USA, pages 362–369.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)*. Philadelphia, USA, pages 311–318.

J. Rebecca Passonneau and Bob Carpenter. 2014. The benefits of a model of annotation. *Transactions of the Association of Computational Linguistics* 2(1):311–326.

Keisuke Sakaguchi, Matt Post, and Benjamin Van Durme. 2014. Efficient elicitation of annotations for human evaluation of machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*. Baltimore, USA, pages 1–11.

Keisuke Sakaguchi and Benjamin Van Durme. 2018. Efficient online scalar annotation with bounded support. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. pages 208–218. <http://aclweb.org/anthology/P18-1020>.

Rion Snow, Brendan O’Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and fast — but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Honolulu, USA, pages 254–263.

## **ALTA Shared Task papers**

# Overview of the 2018 ALTA Shared Task: Classifying Patent Applications

Diego Mollá

Department of Computing  
Macquarie University  
Sydney, Australia

diego.molla-aliod@mq.edu.au

Dilesha Seneviratne

Queensland University of Technology (QUT)  
Brisbane, Australia

d.dwmhakmanawalawwe@hdr.qut.edu.au

## Abstract

We present an overview of the 2018 ALTA shared task. This is the 9th of the series of shared tasks organised by ALTA since 2010. The task was to classify Australian patent classifications following the sections defined by the International Patent Classification (IPC), using data made available by IP Australia. We introduce the task, describe the data and present the results of the participating teams. Some of the participating teams outperformed state of the art.

## 1 Introduction

When a patent application is submitted there is a process where the application is classified by examiners of patent offices or other people. Patent classifications make it feasible to search quickly for documents about earlier disclosures similar to or related to the invention for which a patent is applied for, and to track technological trends in patent applications.

The International Patent Classification (IPC) is a hierarchical patent classification system that has been agreed internationally. The first edition of the classification was established by the World Intellectual Property Organization (WIPO) and was in force from September 1, 1968 ([WIPO, 2018](#)). The classification has undertaken a number of revisions since then. Under the current version, a patent can have several classification symbols but there is one which is the primary one. This is what is called the *primary IPC mark*.

An IPC classification symbol is specified according to a hierarchy of information. The generic form of the symbol is A01B 1/00, where each component has a special meaning as defined by [WIPO \(2018\)](#). The first character of the IPC clas-

## Symbol    Section

A	Human necessities
B	Performing operations, transporting
C	Chemistry, metallurgy
D	Textiles, paper
E	Fixed constructions
F	Mechanical engineering, lighting, heating, weapons, blasting
G	Physics
H	Electricity

Table 1: Sections of the IPC

sification symbol denotes the first level of the hierarchy or *section symbol*. This is a letter from A to H as defined in Table 1.

The goal of the 2018 ALTA Shared Task is to automatically classify Australian patents into one of the IPC sections A to H. Section 2 introduces the ALTA shared tasks. Section 3 presents some related work. Section 4 describes the data. Section 5 describes the evaluation criteria. Section 6 presents the results, and Section 7 concludes this paper.

## 2 The 2018 ALTA Shared Task

The 2018 ALTA Shared Task is the 9th of the shared tasks organised by the Australasian Language Technology Association (ALTA). Like the previous ALTA shared tasks, it is targeted at university students with programming experience, but it is also open to graduates and professionals. The general objective of these shared tasks is to introduce interested people to the sort of problems that are the subject of active research in a field of natural language processing.

There are no limitations on the size of the teams or the means that they can use to solve the problem, as long as the processing is fully automatic

— there should be no human intervention.

As in past ALTA shared tasks, there are two categories: a student category and an open category.

- All the members of teams from the **student category** must be university students. The teams cannot have members that are full-time employed or that have completed a PhD.
- Any other teams fall into the **open category**.

The prize is awarded to the team that performs best on the private test set — a subset of the evaluation data for which participant scores are only revealed at the end of the evaluation period (see Section 5).

### 3 Related Work

Extensive research has been conducted on automating patent classification in the IPC hierarchy and a wide variety of approaches have been proposed. These approaches use features that are generated/extracted from patent content (claim, description, etc), patent metadata (title, applicant name, filing date, inventor name, etc) and citations to represent patent documents in classification (Liu and Shih, 2011). Patent content-based features are the most popular choice among the different types of features to address patent classification (Liu and Shih, 2011). In addition, features based on patent metadata which are considered to have strong classification power have been used to boost the classification performance (Richter and MacFarlane, 2005). Further, patents are not isolated but they are connected through citations which provide rich information about the patent network. Thus, researchers have utilised patent citation information to generate features for patent classification (Liu and Shih, 2011; Li et al., 2007). While all these types of features have served to build classifiers, which features can represent the patents well is still an open question (Gomez and Moens, 2014b).

Some of the widely used classification algorithms in the literature for building patent classification systems are Naive Bayes (NB), Artificial Neural Network (ANN), Support Vector Machines (SVM), K-Nearest Neighbour (KNN), Decision Trees (DT) and Logistic Regression (LR). The greater part of these systems has focused on achieving classification effectiveness. SVM has shown superior performance in terms of effectiveness with some datasets (Fall et al., 2003), yet

it has not been able to scale with large datasets. Seneviratne et al. (2015) have proposed a document signature-based patent classification approach employing KNN which can address the scalability and efficiency with a competitive effectiveness.

Given that there are different evaluation measures and different datasets, it is difficult to compare the performance between many patent classification approaches. Apart from the shared evaluation tasks of patent classification like CLEF-IP 2010 (Piroi et al., 2010) and CLEF-IP 2011 (Piroi et al., 2011), where the performance of systems were evaluated using benchmark datasets, a limited number of approaches — e.g. by Fall et al. (2003), Tikk et al. (2005) and Seneviratne et al. (2015) — have evaluated their methods using publicly available complete data sets like WIPO-alpha<sup>1</sup> and WIPO-de.<sup>2</sup> The majority of other systems have been evaluated using *ad-hoc* datasets, making it difficult to extrapolate their performance (Gomez and Moens, 2014b).

The CLEF-IP 2010 and 2011 classification tasks required to classify patents at the IPC subclass level (Piroi et al., 2010, 2011), which is finer grained than the section level used in the ALTA shared task. Both of these classification tasks used evaluation measures such as Precision@1, Precision@5, Recall@5, Map and F1 at 5, 25 and 50. While the best results of these experiments varied, the best results were from Verberne and D'hondt (2011), who achieved 0.74, 0.86, and 0.71 for precision, recall, and F1 score respectively.

Most of the researchers who have conducted experiments with complete WIPO-alpha and WIPO-de datasets have reported their results at IPC section and subclass levels. For example, the hierarchical classification method by Tikk et al. (2005) has achieved an accuracy of 0.66 at the section level with the WIPO-alpha dataset and 0.65 with the WIPO-de dataset. Gomez and Moens (2014a) have reported their classification results for WIPO-alpha at the section level and the reported values for accuracy and macro-averaged F1 score are 0.74 and 0.71 respectively.

<sup>1</sup> <http://www.wipo.int/classifications/ ipc/en/ITsupport/Categorization/dataset/wipo-alpha-readme.html>

<sup>2</sup> <http://www.wipo.int/classifications/ ipc/en/ITsupport/Categorization/dataset/index.html>

ID	Label
0	A
1	G
2	A
3	A
4	D
5	A

Table 2: First 5 rows of the training data

## 4 Data

The data used in the 2018 ALTA Shared Task consists of a collection of Australian patents partitioned into 3,972 documents for training and 1,000 documents for test. The documents are plain text files which are the result of applying a text extracting tool on the original PDF files. As a result, there are errors in the documents, some of which are documented by the participants of the shared task (Benites et al., 2018; Hepburn, 2018). In particular, 61 documents contain the string “NA[newline]parse failure”. In addition, meta-data information such as titles, authors, etc. are not marked up in the documents.

The data have been anonymised by replacing the original file names with unique IDs starting from number 1. Prior to assigning the IDs, the files have been shuffled and split into the training and test sets. Two CSV files are used to specify the training and test data, so that the training data contains the annotated sections, and the test data only contain the IDs of the test documents. Table 2 shows the first lines of the CSV file specifying the training data.

Figure 1 shows the label distributions of the training and test data. There was no attempt to obtain stratified splits and consequently there were slight differences in the distributions of labels. We can also observe a large imbalance in the distribution of labels, where the most frequent label (“A”) occurs in more than 30% of the data, and the least frequent label (“D”) occurs in only 0.2% to 0.3% of the data.

## 5 Evaluation

As in previous ALTA shared tasks, the 2018 shared task was managed and evaluated using Kaggle in Class, with the name “ALTA 2018 Chal-

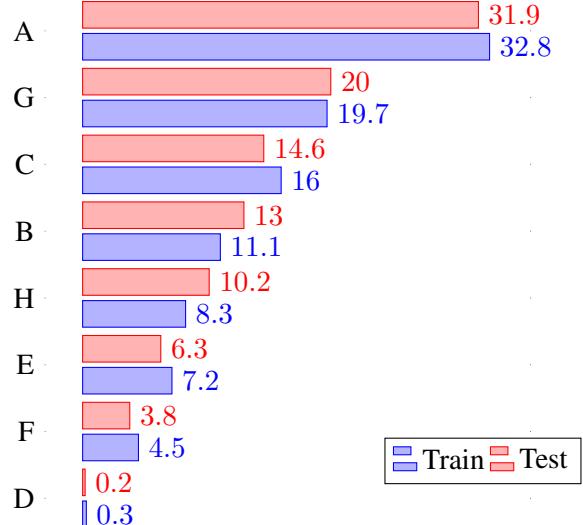


Figure 1: Distribution of labels in percentages

lenge”.<sup>3</sup> This enabled the participants to submit runs prior to the submission deadline for immediate feedback and compare submissions in a leaderboard.

The framework provided by Kaggle in Class allowed the partition of the test data into a public and a private section. Whenever a participating team submitted a run, the evaluation results of the public partition were immediately available to the team, and the best results of each team appeared in the public leaderboard. The evaluation results of the private partition were available to the competition organisers only, and were used for the final ranking after the submission deadline. To split the test data into the public and private partitions, we used the defaults provided by Kaggle in Class. These defaults performed a random partition with 50% of the data falling into the public partition, and the remaining 50% falling into the private partition. The participants were able to see the entire unlabelled evaluation data, but they did not know what part of the evaluation data belonged to which partition.

Each participating team was allowed to submit up to two (2) runs per day. By limiting the number of runs per day, and by not disclosing the results of the private partition, the risks of overfitting to the private test results were controlled.

The chosen evaluation metric was the micro-averaged F1 score. This metric is common in

<sup>3</sup><https://www.kaggle.com/c/alta-2018-challenge>

multi-label classification tasks, and measures the harmonic mean of recall and precision according to the formula:

$$F1 = 2 \frac{p \cdot r}{p + r}$$

Where  $p$  is the precision computed as the ratio of true positives to all predicted positives, and  $r$  is the recall computed as the ratio of true positives to all actual positives. In particular:

$$p = \frac{\sum_{k \in C} tp_k}{\sum_{k \in C} tp_k + \sum_{k \in C} fp_k}$$

$$r = \frac{\sum_{k \in C} tp_k}{\sum_{k \in C} tp_k + \sum_{k \in C} fn_k}$$

Where  $tp_k$ ,  $fp_k$  and  $fn_k$  are the number of true positives, false positives, and false negatives, respectively, in class  $k \in \{A, B, C, D, E, F, G, H\}$ .

## 6 Results

A total of 14 teams registered in the student category, and 3 teams registered in the open category. Due to the nature of the Kaggle in Class framework, Kaggle users could register to the Kaggle system and submit runs without notifying the ALTA organisers, and therefore a number of runs were from unregistered teams. In total, 14 teams submitted runs, of which 6 were registered in the student category and 3 were registered in the open category. The remaining teams were disqualified for the final prize. Table 3 shows the results of the public and private submissions of all teams, including the runs of disqualified teams.

Table 3 also includes two baselines. The Naive Bayes baseline was made available to the participants as a Kaggle kernel.<sup>4</sup> The baseline implemented a simple pipeline using the sklearn environment<sup>5</sup> that implemented a Naive Bayes classifier using  $tf.idf$  features. Both the Naive Bayes classifier and the  $tf.idf$  vectoriser used the defaults provided by sklearn and were not fine-tuned. All of the participant's best runs outperformed the baseline.

The SIG\_CLS baseline is the system reported by Seneviratne et al. (2015). The system was retrained with the shared task data with small

Team	Category	Private	Public
BMZ	Open	<b>0.778</b>	0.776
Jason Hepburn	Student	0.764	<b>0.784</b>
Forefront Analytics	Open	0.732	0.722
(disqualified)	—	0.722	0.704
NLPGirls	Student	0.702	0.748
Western Journalists	Student	0.702	0.742
ANUCompGrads	Student	0.698	0.720
NLP-CIC	Student	0.696	0.712
Hemu	Student	0.694	0.726
SIG_CLS	baseline	0.650	0.638
HAL9000	Open	0.630	0.646
(disqualified)	—	0.626	0.656
(disqualified)	—	0.604	0.638
Naive Bayes	baseline	0.408	0.448

Table 3: Micro-averaged F1 of the best public and private runs

changes on the system settings.<sup>6</sup> Virtually all participants obtained better results than this second baseline.

In past competitions of the ALTA shared task there were some differences between the rankings given in the public and the private submissions. This is the first time, however, that the best teams in the public and the private runs differ. Following the rules of the shared task, the winning team was BMZ, and the best team in the student category was Jason Hepburn. These two teams describe their system in separate papers (Benites et al., 2018; Hepburn, 2018).

## 7 Conclusions

The 2018 ALTA Shared Task was the 9th of the series of shared tasks organised by ALTA. This year's task focused on document classification of Australian patent applications following the sections defined by the International Patent Classification (IPC). There was very active participation, with some teams submitting up to 30 runs. Participation was increasingly active near the final submission date, and the top rows of the public leaderboard changed constantly. To the best of our knowledge, prior to this shared task the best-performing system using the WIPO-alpha set reported an accuracy of 0.74 and a macro-averaged F1 score of 0.71 (Gomez and Moens, 2014a). Ta-

<sup>4</sup><https://www.kaggle.com/dmollaaliod/naive-bayes-baseline>

<sup>5</sup><https://scikit-learn.org/stable/>

<sup>6</sup>The specific system settings were: signature width of 8,192 bits, and 10-nearest neighbours. The complete patent text was used to build the patent signatures.

Team	Test Data	Micro-F1	Macro-F1	Accuracy
BMZ	ALTA	<b>0.78</b>	<b>0.75</b>	<b>0.78</b>
Jason Hepburn	ALTA	0.77	<b>0.75</b>	0.77
Gomez and Moens (2014a)	WIPO-alpha		0.71	0.74
Tikk et al. (2005)	WIPO-alpha			0.66

Table 4: Micro-F1, Macro-F1 and Accuracy of best-performing systems and comparison with literature.

ble 4 shows the accuracy and micro- and macro-averaged F1 score of the two top-performing systems in the test set of the ALTA shared task.<sup>7</sup> Both systems achieved better results in all comparable metrics, which indicates that they appear to have outperformed the state of the art.

## Acknowledgments

This shared task was made possible thanks to the data provided by the Digital Transformation Agency and IP Australia.

## References

- Fernando Benites, Shervin Malmasi, and Marcos Zampieri. 2018. Classifying patent applications with ensemble methods. In *Proceedings ALTA 2018*.
- Caspar J Fall, Atilla Törcsvári, Karim Benzineb, and Gabor Karetka. 2003. Automated categorization in the international patent classification. In *AcM Sigir Forum*. ACM, volume 37, pages 10–25.
- Juan Carlos Gomez and Marie-Francine Moens. 2014a. Minimizer of the reconstruction error for multi-class document categorization. *Expert Systems with Applications* 41(3):861–868.
- Juan Carlos Gomez and Marie-Francine Moens. 2014b. A survey of automated hierarchical classification of patents. In *Professional Search in the Modern World*, Springer, pages 215–249.
- Jason Hepburn. 2018. Universal language model fine-tuning for patent classification. In *Proceedings ALTA 2018*.
- Xin Li, Hsinchun Chen, Zhu Zhang, and Jiexun Li. 2007. Automatic patent classification using citation network information: an experimental study in nanotechnology. In *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*. ACM, pages 419–427.
- Duen-Ren Liu and Meng-Jung Shih. 2011. Hybrid-patent classification based on patent-network analysis. *Journal of the American Society for Information Science and Technology* 62(2):246–256.
- Florina Piroi, Mihai Lupu, Allan Hanbury, Alan P Sexton, Walid Magdy, and Igor V Filippov. 2010. Clef-ip 2010: Retrieval experiments in the intellectual property domain. In *CLEF (notebook papers/labs/workshops)*.
- Florina Piroi, Mihai Lupu, Allan Hanbury, and Veronika Zenz. 2011. Clef-ip 2011: Retrieval in the intellectual property domain. In *CLEF (notebook papers/labs/workshop)*.
- Georg Richter and Andrew MacFarlane. 2005. The impact of metadata on the accuracy of automated patent classification. *World Patent Information* 27(1):13–26.
- Dilesha Seneviratne, Shlomo Geva, Guido Zuccon, Gabriela Ferraro, Timothy Chappell, and Magali Meireles. 2015. A signature approach to patent classification. In *Asia Information Retrieval Symposium*. Springer, pages 413–419.
- Domonkos Tikk, György Biró, and Jae Dong Yang. 2005. Experiment with a hierarchical text categorization method on wipo patent collections. In *Applied Research in Uncertainty Modeling and Analysis*, Springer, pages 283–302.
- Suzan Verberne and Eva D’hondt. 2011. Patent classification experiments with the linguistic classification system lcs in clef-ip 2011. In *CLEF (Notebook Papers/Labs/Workshop)*.
- WIPO. 2018. Guide to the international patent classification, version 2018. Technical report, World Intellectual Property Organization.

<sup>7</sup>Due to a glitch with the Kaggle framework we were unable to determine the public and private partitions of the test set. Consequently, the systems were tested on the *combined* public and private partitions.

# Classifying Patent Applications with Ensemble Methods

Fernando Benites<sup>1</sup>, Shervin Malmasi<sup>2,3</sup>, Marcos Zampieri<sup>4</sup>

<sup>1</sup>Zurich University of Applied Sciences, Switzerland

<sup>2</sup>Harvard Medical School, United States

<sup>3</sup>Macquarie University, Australia

<sup>4</sup>University of Wolverhampton, United Kingdom

benf@zhaw.ch, shervin.malmasi@mq.edu.au, m.zampieri@wlv.ac.uk

## Abstract

We present methods for the automatic classification of patent applications using an annotated dataset provided by the organizers of the ALTA 2018 shared task - Classifying Patent Applications. The goal of the task is to use computational methods to categorize patent applications according to a coarse-grained taxonomy of eight classes based on the International Patent Classification (IPC). We tested a variety of approaches for this task and the best results, 0.778 micro-averaged F1-Score, were achieved by SVM ensembles using a combination of words and characters as features. Our team, BMZ, was ranked first among 14 teams in the competition.

## 1 Introduction

According to statistics of the World Intellectual Property Organization (WIPO),<sup>1</sup> the number of patent applications filled across the world keeps growing every year. To cope with the large volume of applications, companies and organizations have been investing in the development of software to process, store, and categorize patent applications with minimum human intervention.

An important part of patent application forms is, of course, composed of text. This has led to the widespread use of NLP methods in patent application processing systems as evidenced in Section 2. One such example is the use of text classification methods to categorize patent applications according to standardized taxonomies such as the International Patent Classification (IPC)<sup>2</sup> as discussed in the studies by Benzineb and Guyot (2011); Fall et al. (2003).

In this paper, we present a system to automatically categorize patent applications from Australia according to the top sections of the IPC taxonomy using a dataset provided by the organizers of the ALTA 2018 shared task on Classifying Patent Applications (Molla and Seneviratne, 2018).<sup>3</sup> The dataset and the taxonomy are presented in more detail in Section 3. Building on our previous work (Malmasi et al., 2016a; Malmasi and Zampieri, 2017), our system is based on SVM ensembles and it achieved the highest performance of the competition.

## 2 Related Work

There have been a number of studies applying NLP and Information Retrieval (IR) methods to patent applications specifically, and to legal texts in general, published in the last few years.

Applications of NLP and IR to legal texts include the use of text summarization methods (Farzindar and Lapalme, 2004) to summarize legal documents and most recently, court ruling prediction. A few papers have been published on this topic, such as the one by Katz et al. (2014) which reported 70% accuracy in predicting decisions of the US Supreme Court, Aletras et al. (2016); Medvedeva et al. (2018) which explored computational methods to predict decisions of the European Court of Human Rights (ECRH), and (Sulea et al., 2017a,b) on predicting the decisions of the French Supreme Court. In addition to the aforementioned studies, one recent shared task has been organized on court rule prediction (Zhong et al., 2018).

Regarding the classification of patent applications, the task described in this paper, a related dataset WIPO-alpha was used in the experiments

<sup>1</sup><http://www.wipo.int/ipstats/en/>

<sup>2</sup><http://www.wipo.int/classifications/ ipc/en/>

<sup>3</sup><http://www.alta.asn.au/events/sharedtask2018/>

and it is often used in such studies. The WIPO-alpha consists of a different number of patents (in the thousands, but it grows every year) and is usually used in its hierarchical call form ([Tikk and Biró, 2003](#)). Recently, word embeddings and LSTMs were applied to the task ([Grawe et al., 2017](#)). There, the experiments were hierarchically conducted but in a superficial manner.

Hoffmann et al. investigated in depth the hierarchical problem of WIPO-alpha with SVMs ([Hoffmann et al., 2003](#); [Tsachantaridis et al., 2004](#); [Cai and Hofmann, 2007](#)). They showed that using a hierarchical approach produced better results. Many studies showed that evaluating a hierarchical classification task is not trivial and many measures can integrate the class ontology. Still, using multiple hierarchical measures can introduce bias ([Brucker et al., 2011](#)). Yet, there was much improvement in the last 3-4 years in the text classification field. This is one reason, why, when reengaging again in the WIPO-alpha dataset, investigating only the top nodes of WIPO class ontology might be a good start for future successive tasks.

Finally, at the intersection between patent applications and legal texts in general, [Wongchaisuwat et al. \(2016\)](#) presented experiments on predicting patent litigation and time to litigation.

### 3 Data

The dataset released by the organizers of the ALTA 2018 shared task consists of a collection of Australian patent applications. The dataset contains 5,000 documents released for training and 1,000 documents for testing. The classes relevant for the task consisted of eight different main branches of the WIPO class ontology as follows:

- A: Human necessities;
- B: Performing operations, transporting;
- C: Chemistry, metallurgy;
- D: Textiles, paper;
- E: Fixed constructions;
- F: Mechanical engineering, lighting, heating, weapons, blasting;
- G: Physics;
- H: Electricity.

The documents were created using automated OCR and therefore, not thoroughly cleaned before release. For example, there were documents

with expressions such as “NA\\nparse failure” and page numbers in the middle of paragraphs which made processing more challenging. We enhanced the dataset with data from the WIPO-alpha repository gathered in October 2018 consisting of 46,319 training documents and 28,924 test documents. We also took a random sub-sample of 100,000 documents from the WIPO-en gamma English dataset, which contains 1.1 million patent documents in total.

We utilized all of the available text fields in the texts and concatenated them into a single document.

## 4 Methodology

### 4.1 Preprocessing

The documents come from different sources and authors, therefore no standard representation exists and there is high variation in formatting across the documents. Since we do not utilize document structure in our approach, we decided to eliminate it by collapsing the documents into a single block of text. This was done by replacing all consecutive non-alphanumeric characters with a single space. Next, we converted the text to lowercase and removed any tokens representing numbers.

### 4.2 Features

For feature extraction we used and extended the methods reported in [Malmasi and Zampieri \(2017\)](#). Term Frequency (TF) of  $n$ -grams with  $n$  ranging from 3 to 6 for characters and 1-2 for words have been used. Along with term frequency we calculated the inverse document frequency (TF-IDF) ([Gebre et al., 2013](#)) which resulted in the best single feature set for prediction.

### 4.3 Classifier

We used an ensemble-based classifier for this task. Our base classifiers are linear Support Vector Machines (SVM). SVMs have proven to deliver very good performance in a number of text classification problems. It was previously used for complex word identification ([Malmasi et al., 2016a](#)), triage of forum posts ([Malmasi et al., 2016b](#)), dialect identification ([Malmasi and Zampieri, 2017](#)), hate speech detection ([Malmasi and Zampieri, 2018](#)), and court ruling prediction ([Sulea et al., 2017a](#)).

	<b>Training</b>	<b>Public (Validation)</b>	<b>Private (Test)</b>
(1) Baseline 20k feats.	0.709	0.710	0.692
(2) Baseline 40k feats.	0.715	-	-
(3) Baseline w/ WIPO-alpha	0.775	0.758	0.744
(4) Semi-supervised	0.734	0.728	0.704
(5) Ensemble w/ WIPO-alpha + gamma	0.787	0.776	0.778

Table 1: F1-micro performance of the systems in training (10-fold CV), in the validation and in the test sets (train, public and private leaderboard).

#### 4.4 Systems

We developed a number of different systems. As baselines we employed single SVM models with TF-IDF, using the top 20k and 40k more frequent words as features, resulting in two models. We created a third baseline which included the WIPO-alpha data for training.

For system 4, we augmented system 3 with a semi-supervised learning approach similar to the submission by Jauhainen et al. (2018) to the dialect identification tasks at the VarDial workshop (Zampieri et al., 2018). This approach consists of classifying the unlabelled test set with a model based on the training data, then selecting the predictions with the highest confidence and using them as new additional training samples. This approach can be very useful if there are few training samples and out-of-domain data is expected.

Finally, for system 5, we extended system 4 to be an ensemble of both word- and character-based models, and to include additional training data from the WIPO-alpha and WIPO-en gamma datasets, as described in 3.

### 5 Results

In this section, we investigate the impact of the different systems and data. We give special attention to the competition results showing these in different settings. This is particularly interesting since the amount of data with WIPO-alpha and the vocabulary of the ALTA data without pre-processing was relatively large.

#### 5.1 Official Results

We present the results obtained in the training stage, the public leaderboard, and the private leaderboard in Table 4.1. The shared task was organized using Kaggle<sup>4</sup>, a data science platform, in which the terms Public Leaderboard and Private

Leaderboard are used referring to what is commonly understood as development or validation phase and test phase. This is important in the system development stage as it helps preventing systems from overfitting. We used 10-fold cross validation in the training setup.

As can be seen in Table 4.1, the ensemble system with additional data achieved the best performance. This can be attributed to the use of large amounts of additional training data, a semi-supervised approach, and an ensemble model with many features.

### 6 Conclusion and Future Work

This paper presented an approach to categorizing patent applications in eight classes of the WIPO class taxonomy. Our system competed in the ALTA 2018 - Classifying Patent Applications shared task under the team name BMZ. Our best system is based on an ensemble of SVM classifiers trained on words and characters. It achieved 0.778 micro-averaged F1-Score and ranked first place in the competition among 14 teams.

We observed that expanding the training data using the WIPO datasets brought substantial performance improvement. This dataset is similar to that provided by the shared task organizers in terms of genre and topics and it contains 15 times more samples. The use of an ensemble-based approach prevented the system from overfitting and providing more robust predictions.

In future work we would like to use hierarchical approaches to classify patent applications using a more fine-grained taxonomy. Finally, we would also like to investigate the performance of deep learning methods for this task.

### Acknowledgments

We would like to thank the ALTA 2018 shared task organizers for organizing this interesting shared task and for replying promptly to our inquiries.

<sup>4</sup><https://www.kaggle.com/>

## References

- Nikolaos Aletras, Dimitrios Tsarapatsanis, Daniel Preoțiuc-Pietro, and Vasileios Lampos. 2016. Predicting Judicial Decisions of the European Court of Human Rights: A Natural Language Processing Perspective. *PeerJ Computer Science* 2:e93.
- Karim Benzineb and Jacques Guyot. 2011. Automated Patent Classification. In *Current challenges in patent information retrieval*, Springer, pages 239–261.
- Florian Brucker, Fernando Benites, and Elena P. Sapozhnikova. 2011. An Empirical Comparison of Flat and Hierarchical Performance Measures for Multi-Label Classification with Hierarchy Extraction. In *Proceedings of KES Part I*.
- Lijuan Cai and Thomas Hofmann. 2007. Exploiting known taxonomies in learning overlapping concepts. In *IJCAI*, volume 7, pages 708–713.
- Casper J Fall, Atilla Törcsvári, Karim Benzineb, and Gabor Karetka. 2003. Automated Categorization in the International Patent Classification. In *AcM Sigir Forum*. ACM, volume 37, pages 10–25.
- Atefeh Farzindar and Guy Lapalme. 2004. Legal Text Summarization by Exploration of the Thematic Structures and Argumentative Roles. *Proceedings of the Text Summarization Branches Out Workshop*.
- Binyam Gebrekidan Gebre, Marcos Zampieri, Peter Wittenburg, and Tom Heskes. 2013. Improving Native Language Identification with TF-IDF Weighting. In *Proceedings of the BEA Workshop*.
- M. F. Grawe, C. A. Martins, and A. G. Bonfante. 2017. Automated Patent Classification Using Word Embedding. In *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*. pages 408–411.
- Thomas Hofmann, Lijuan Cai, and Massimiliano Ciaramita. 2003. Learning with taxonomies: Classifying documents and words. In *NIPS workshop on syntax, semantics, and statistics*.
- Tommi Jauhainen, Heidi Jauhainen, and Krister Lindén. 2018. Heli-based experiments in swiss german dialect identification. In *Proceedings of the VarDial Workshop*.
- Daniel Martin Katz, Michael J. Bommarito II, and Josh Blackman. 2014. Predicting the behavior of the supreme court of the united states: A general approach. *CoRR* abs/1407.6333.
- Shervin Malmasi, Mark Dras, and Marcos Zampieri. 2016a. LTG at SemEval-2016 task 11: Complex Word Identification with Classifier Ensembles. In *Proceedings of SemEval*.
- Shervin Malmasi and Marcos Zampieri. 2017. German dialect identification in interview transcriptions. In *Proceedings of the VarDial Workshop*.
- Shervin Malmasi and Marcos Zampieri. 2018. Challenges in Discriminating Profanity from Hate Speech. *Journal of Experimental & Theoretical Artificial Intelligence* 30(2):187–202.
- Shervin Malmasi, Marcos Zampieri, and Mark Dras. 2016b. Predicting Post Severity in Mental Health Forums. In *Proceedings of CLPsych Workshop*.
- Masha Medvedeva, Michel Vols, and Martijn Wieling. 2018. Judicial Decisions of the European Court of Human Rights: Looking into the Crystal Ball. *Proceedings of the Conference on Empirical Legal Studies*.
- Diego Molla and Dilesha Seneviratne. 2018. Overview of the 2018 ALTA Shared Task: Classifying Patent Applications. In *Proceedings of ALTA*.
- Octavia-Maria Sulea, Marcos Zampieri, Shervin Malmasi, Mihaela Vela, Liviu P Dinu, and Josef van Genabith. 2017a. Exploring the use of text classification in the legal domain. *arXiv preprint arXiv:1710.09306*.
- Octavia-Maria Sulea, Marcos Zampieri, Mihaela Vela, and Josef van Genabith. 2017b. Predicting the Law Area and Decisions of French Supreme Court Cases. In *Proceedings of RANLP*.
- Domonkos Tikk and György Biró. 2003. Experiment with a hierarchical text categorization method on the wipo-alpha patent collection. In *Proceedings of ISUMA 2003*.
- Ioannis Tsochantaridis, Thomas Hofmann, Thorsten Joachims, and Yasemin Altun. 2004. Support vector machine learning for interdependent and structured output spaces. In *Proceedings of the ICML*.
- Papis Wongchaisuwat, Diego Klabjan, and John O McGinnis. 2016. Predicting Litigation Likelihood and Time to Litigation for Patents. *arXiv preprint arXiv:1603.07394*.
- Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Ahmed Ali, Suwon Shon, James Glass, Yves Scherer, Tanja Samardžić, Nikola Ljubešić, Jörg Tielemann, et al. 2018. Language Identification and Morphosyntactic Tagging: The Second VarDial Evaluation Campaign. In *Proceedings of VarDial Workshop*.
- Haoxi Zhong, Chaojun Xiao, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, et al. 2018. Overview of CAIL2018: Legal Judgment Prediction Competition. *arXiv preprint arXiv:1810.05851*.

# Universal Language Model Fine-tuning for Patent Classification

Jason Hepburn

Macquarie University

Sydney, Australia

jason.hepburn@students.mq.edu.au

## Abstract

This paper describes the methods used for the 2018 ALTA Shared Task. The task this year was to automatically classify Australian patents into their main International Patent Classification section. Our final submission used a Support Vector Machine (SVM) and Universal Language Model with Fine-tuning (ULMFiT). Our system achieved the best results in the student category.

## 1 Introduction

For the last nine years the Australasian Language Technology Association (ALTA) has run a shared task competition for students. This year the shared task is to classify patent applications into their primary section code ([Mollá and Seneviratne, 2018](#)).

Patent applications are classified and compared to previous inventions in the field. Accurate classification of patents is crucial to patent officers, potential inventors, and industry. The patent classification process is dependant on human labour and with the rate of submissions increasing there is an ever greater need for an Automated Patent Classification (APC) system ([Fall et al., 2003](#)).

The International Patent Classification (IPC) has a tree structured class hierarchy ([Silla and Freitas, 2011](#)). At the highest level of this hierarchy is the IPC Section designated by the capital letters A to H (Table 1). Following the tree structure from Sections are Classes, Sub-classes, and Groups. There are approximately 69,000 different categories at the group level. The classification taxonomy is revised annually and previous patents can be reclassified ([D'hondt et al., 2013](#)).

Most patents have a main code in addition to a set of secondary codes. These secondary codes can be very distant to each other. For some codes

- A Human necessities
- B Performing operations, transporting
- C Chemistry, metallurgy
- D Textiles, paper
- E Fixed constructions
- F Mechanical engineering, lighting, heating, weapons, blasting
- G Physics
- H Electricity

Table 1: IPC Sections

it is obligatory to also assign other codes (eg. All C12N are also classed A61P). Codes can have *placement* rules defining a preference for one code when two may apply.

At the semantic level all patents are different as they must describe a new idea or invention. Some terms, phrases, or acronyms can have very different meaning in different fields. Applicants try to avoid narrowing the scope of the invention and as such can use vague or general terms. As an example, pharmaceutical companies tend to describe every possible therapeutic use for an application. This can make it difficult to classify these patents.

We structure this paper as follows: Section 2 introduces related research for APC; Section 3 describes the data set provided for the competition; Section 4 describes the methods used; Section 5 presents and discusses the results; Section 6 concludes this paper.

## 2 Related works

With the need for reliable and efficient APC systems considerable research has been conducted in this area.

[Fall et al. \(2003\)](#) introduce the publicly available WIPO-alpha data set for patent classification (See section 3.2). They give a comprehensive description of the problem and much of its complex-

ties. One such complexity is the similarities of section G and H which are "Physics" and "Electricity" respectively. The authors give a detailed analysis of the classification errors between these two sections.

Various classification models are tested and compared including Naïve Bayes, K-Nearest Neighbours, and SVM. Fall et al. (2003) show that the best performing model is a SVM with a linear kernel using only the first 300 words of the document.

Benzineb and Guyot (2011) describe in great detail the task and challenges of APC. APC can be used to classify new applications as well as help with searches for similar prior art. Interestingly they noted that SVMs are more accurate than Neural Network approaches.

D'hondt et al. (2013) assess the use of statistical and linguistic phrases for patent applications. Adding phrases, particularly bigrams, to unigrams significantly improves classification.

Seneviratne et al. (2015) build on Falls work with a focus on improving the efficiency of classification. Dimensionality reduction is used in the form of a signature approach to reduce computation and enable a larger vocabulary. For top predictions a marginal improvement is made.

### 3 Data sets

In this section we describe the two data sets used by our system. The first data set is provided for the ALTA Shared task<sup>1</sup>. The second is the WIPO-alpha data set introduced by Fall et al. (2003).

#### 3.1 ALTA

The data provided contains 4972 Australian patent applications. 3972 of them are part of the training set labelled with the main IPC section. The other 1000 applications in the test set are unlabelled.

The section counts are significantly unbalanced with the largest, section A, having 1303 compared to section D having 14 (see Figure 1).

#### 3.2 WIPO-alpha

WIPO-alpha is a collection of patent applications from the World Intellectual Property Organization. The documents are all in English and published between 1998 and 2002. Each patent is a structured XML document. This allows for analysis of separate parts of the documents such as the title

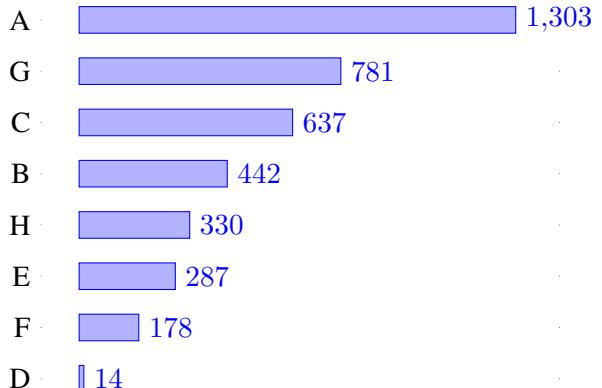


Figure 1: ALTA training set counts by IPC Section

or abstract. Documents include the full IPC main classification as well as secondary classifications.

There are 75,250 documents in the data set split into approximately 60% train and 40% test. The splitting of the train and test sets has tried to maintain an equal distribution IPC main group level.

## 4 Methodology

We used several statistical classifiers to complete this task. In this section we describe in detail the methods used and the steps they involved. Section 4.1 describes the pre-processing of the ALTA and WIPO-alpha data sets. Section 4.2 describes the SVM classifier motivated by Fall et al. (2003). Section 4.3 describes ULMFiT from Howard and Ruder (2018) and how it is adapted to this task. Section 4.5 describes the system used to deal with the classification errors between Section G and H.

#### 4.1 Pre-processing

During the exploration of the data it was found that there is a large variation of document length. There are 48 documents in the ALTA training data set which contained only "NA parse failure". These documents were excluded from the training set and when found in the test set automatically classified as Section A which is the majority class. Looking closer at the large documents some contain long strings of DNA and amino acid sequences. The largest document appears to contain a large number of incorrectly encoded characters. Motivated by Fall et al. (2003), this and other noisy data is avoided by only using a small portion of the beginning of the document.

<sup>1</sup>[www.alta.asn.au/events/sharedtask2018](http://www.alta.asn.au/events/sharedtask2018)

Patent documents from the WIPO data set are in XML format. These documents were converted into plain text to best replicate the format of the target ALTA documents. This was achieved by concatenating the document Title, Abstracts, and Claim.

## 4.2 SVM

For the SVM classifier we use the Python Scikit-learn (Pedregosa et al., 2011) library. Documents are indexed using term frequency-inverse document frequency (tf-idf) and only using the first 3500<sup>2</sup> characters. Motivated by D’hondt et al. (2013) we use unigrams and bigrams. As with the work of Fall et al. (2003) linear kernels for the SVM were found to perform best.

## 4.3 ULMFiT

Universal Language Model Fine-tuning (ULMFiT) is a transfer learning technique introduced by Howard and Ruder (2018). This technique uses the following three steps: a) General-domain language model pretrainig (4.3.1); b) Target task language model fine-tuning(4.3.2); and c) Target task classifier fine-tuning (4.3.3).

### 4.3.1 General-domain language model pretraining

The first step is to carry out unsupervised training of a language model on a large corpus to create a general-domain language model. As this step is not domain specific here we have used the pretrained model<sup>3</sup> from Howard and Ruder (2018). This model uses the state of the art language model AWD LSTM trained on Wikitext-103 (Merity et al., 2017)

### 4.3.2 Target task language model fine-tuning

The general-domain language model is then fine-tuned on data from the target task. The pretraining allows this stage to converge faster and results in a robust language model even for small datasets. A key advantage here is that words that are uncommon in the target training set retain robust representations from the pretraining. As this fine-tuning is also unsupervised here we use both the ALTA training and test sets as well as the WIPO-alpha training set <sup>4</sup>.

<sup>2</sup>Testing of different lengths found that 3500 characters performed best.

<sup>3</sup><http://files.fast.ai/models/wt103/>

<sup>4</sup>Fine-tuning on only the ALTA data set performed poorly compared to SVM

Data	Model	Private	Public	Mean
ALTA	SVM	0.714	0.722	0.718
	ULMFiT	0.662	0.712	0.687
WIPO	SVM	0.684	0.728	0.706
	ULMFiT	0.738	0.730	0.734
Both	SVM	0.748	0.754	0.751
	ULMFiT	<b>0.770</b>	0.760	0.765
Ensemble		0.764	0.772	0.768
Ensemble + G/H		0.752	<b>0.784</b>	0.768

Table 2: F1 scores

### 4.3.3 Target task classifier fine-tuning

The final step adds two additional linear blocks to the pretrained language model. The first linear layer takes as the input the pooled last hidden layers of the language model and applies a ReLU activation. The last layer is fully connected with a softmax activation to output the probability over the target classes.

## 4.4 Ensemble

The ensemble stage is combined using hard voting. The four systems that had the highest results on the public set were used. Specifically this includes SVM and ULMFiT trained only with WIPO-alpha and the same models trained with the combined ALTA and WIPO-alpha data. Ties were broken by defaulting to the best performing system which was ULMFiT trained on the combined ALTA and WIPO-alpha data.

## 4.5 G/H decider

To reduce many of the errors that occur between section G and H we use two more SVM classifiers trained only on the ALTA training set. The first is a binary classifier to separate the G/H from Not G/H. The second classifier is trained to separate section G from H. These classifiers were applied at the ensemble stage such that if the first model classified the document as G/H then the ensemble label was overridden by the G or H label of the second model.

## 5 Results

Results for this task were evaluated by micro-averaged F1-Score and shown in Table 2.

When only using the smaller ALTA data set SVM outperformed ULMFiT. Training with the larger WIPO-alpha data significantly improved the performance of ULMFiT. This validated the use of

the WIPO-alpha data set as it performed better on the ALTA test set despite not using the ALTA data for training.

Training with both data sets together improved both models further.

The performance of some models turned out to be quite different on the private and public splits of the test set. The model that performed best on the public set was third on the private set and the best performance on the private set was third on the public set. The final results on the Kaggle<sup>5</sup> leaderboard also showed similar changes in results for other teams.

Kaggle’s default is to take the two best performing submissions from the public scores as the final submission to the competition. From these two the best private score is used as the final result. This mean that our best performing private score was not available for the final result.

When viewing only the public results it appeared that the Ensemble with G/H decider (section 4.5) performed best. The mean of the public and private scores show that both ensembles performed the same with a score of 0.768. The best private score was achieved with ULMFiT trained on both the ALTA and WIPO-alpha data.

## 6 Conclusion

Patent classification for the 2018 ALTA Shared Task has proven to be a good representation of the challenges of Language Technology. In this paper we describe some of the challenges of patent classification. We show that ULMFiT outperforms SVM for patent classification.

## Acknowledgments

We would like to thank Dr. Diego Mollá Aliod for his time and support with this task and paper.

## References

- Karim Benzineb and Jacques Guyot. 2011. *Automated Patent Classification*, Springer Berlin Heidelberg, Berlin, Heidelberg, pages 239–261. [https://doi.org/10.1007/978-3-642-19231-9\\_12](https://doi.org/10.1007/978-3-642-19231-9_12).
- Eva D’hondt, Suzan Verberne, Cornelis Koster, and Lou Boves. 2013. Text representations for patent classification. *Computational Linguistics* 39(3):755–775. [https://doi.org/10.1162/COLI\\_a\\_00149](https://doi.org/10.1162/COLI_a_00149).

C. J. Fall, A. Törcsvári, K. Benzineb, and G. Karetka. 2003. Automated categorization in the international patent classification. *SIGIR Forum* 37(1):10–25. <https://doi.org/10.1145/945546.945547>.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pages 328–339. <http://aclweb.org/anthology/P18-1031>.

Stephen Merity, Nitish Shirish Keskar, and Richard Socher. 2017. Regularizing and Optimizing LSTM Language Models. *arXiv preprint arXiv:1708.02182*.

Diego Mollá and Dilesha Seneviratne. 2018. Overview of the 2018 alta shared task: Classifying patent applications. In *Proceedings 2018 Australasian Language Technology Workshop ALTA 2018*.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Pas-  
sos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

Dilesha Seneviratne, Shlomo Geva, Guido Zuccon, Gabriela Ferraro, Timothy Chappell, and Magali Meireles. 2015. A signature approach to patent classification. In Guido Zuccon, Shlomo Geva, Hideo Joho, Falk Scholer, Aixin Sun, and Peng Zhang, editors, *Information Retrieval Technology*. Springer International Publishing, Cham, pages 413–419.

Carlos N. Silla and Alex A. Freitas. 2011. A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery* 22(1):31–72. <https://doi.org/10.1007/s10618-010-0175-9>.

<sup>5</sup>[www.kaggle.com/c/alta-2018-challenge](http://www.kaggle.com/c/alta-2018-challenge)