ERC Consolidator Grant 2015
Research Proposal [Part B1]

Translating from Multiple Modalities into Text

TransModal

Principal Investigator: Maria Lapata
Host Institution: University of Edinburgh
Project Duration: 60 months

Proposal Summary
Recent years have witnessed the development of a wide range of computational methods and tools that process and generate natural language text. Many of these have become familiar to mainstream computer users such as tools that retrieve documents matching a query, perform sentiment analysis, and translate between languages. Indeed, publicly available systems like Google Translate can instantly translate between any pair of over fifty human languages allowing users to access web content that wouldn't have otherwise been available. The accessibility of the web could be further enhanced with applications that not only translate between *different languages* (e.g., from English to French) but also *within the same language*, between *different modalities*, or *different data formats*. For example, there are currently no standard tools for simplifying language, e.g., for low-literacy readers or second language learners. The web is rife with non-linguistic data (e.g., databases, graphs, images, mathematical formulas, source code) that cannot be indexed or searched since most retrieval tools operate over textual data.

In this project we argue that in order to render electronic data more accessible to individuals and computers alike, new types of translation models need to be developed. Our proposal is to provide a unified modeling framework for translating from comparable corpora, i.e., collections consisting of data in the same or different modalities that address the same topic without being direct translations of each other. Our objective is to develop a general and scalable modeling framework that can solve different translation tasks and learn the necessary intermediate representations of the units involved in an unsupervised manner without extensive human involvement. We will take advantage of recent advances in *deep learning* to induce general representations for different modalities and learn how these are rendered in natural language. We will follow the general paradigm of *encoder-decoder* modeling where an *encoder* transforms the input into a representation and a *decoder* reconstructs the input and in our case produces the corresponding translation. Advantageously, the encoder-decoder architecture can be jointly trained to maximize the probability of the output given the input, without having to a priori decide about the units of the translation and their interactions. Beyond addressing a fundamental aspect of the translation problem, the proposed framework will lead to the development of novel internet-based applications that automatically simplify text, produce documentation for source code, index images with meaningful descriptions, and summarize database content in natural language.

a. Extended Synopsis

1. State of the Art and Objectives

Natural Language Processing (NLP) is a field of computer science and artificial intelligence concerned with enabling computers to analyze and generate natural language text. Many NLP tools have nowadays reached mainstream computer users. Examples include information management tools that retrieve documents matching a query, perform sentiment analysis, and notably statistical machine translation (SMT). SMT is characterized by the use of machine learning methods and large amounts of previously translated text, known as a parallel corpus or bitext. With an SMT toolkit and enough parallel text, one can build a translation system for a new language pair within as little as a day (Oard and Och, 2003). Publicly available systems like Google Translate can instantly translate between any pair of over fifty human languages and thus allow users to access web content that otherwise would have been unavailable.

The accessibility of the web could be further enhanced with applications that not only translate between *different languages* (e.g., from English to French) but also *within the same language*, and between *different modalities*. Examples include tools that simplify language, e.g., for low-literacy readers, children, or second language learners, extract highlights from documents or create textual descriptions for non-linguistic data such as databases, mathematical formulas, source code, or images. Indeed, the web is rife with non-textual data that cannot be indexed or searched. The ability to translate this information into textual output would enable users to access it and NLP tools to process it. So why haven't these translation problems met with the success of SMT?

To begin with, there are no naturally occurring parallel corpora to learn from. Consider the simplification problem mentioned above; one would need a parallel corpus consisting of low readability[1] texts and their simpler variants. Such monolingual corpora are scarce or even non-existent compared to the abundance of bilingual parallel data and expensive to produce in quantities sufficient for building a good translation system. It is possible to find non-linguistic data accompanied with verbalizations for some of their content. For example, images are often embedded in documents that describe the objects or events depicted in them. Databases (e.g., weather statistics or stock market data) are sometimes collocated with textual descriptions (e.g., weather forecasts, stock market summaries). Source code is accompanied with natural language specifications (e.g., descriptions of a new library, what it does and how to use it). However, the non-linguistic data and the collateral text do not together constitute a clean parallel corpus, but rather a noisy *comparable* corpus.

Secondly, the translation task itself is inherently different. When translating between two languages, it is assumed that the translation process preserves the content being communicated. However, the description of an image does not refer to every single object shown. Analogously, when simplifying a document one may omit certain concepts or elaborate on others. This *content selection* process dictates a modeling approach different from standard SMT learning methods.

In this project we maintain that in order to render electronic data more accessible to individuals and computers alike, new types of translation models need to be developed. Our proposal is to provide a unified modeling framework for translating from comparable corpora, i.e., collections consisting of data in the same or different modalities that address the same topic without being translations of each other. Our key insight is to develop general and scalable models that can solve different translation tasks and learn the necessary intermediate representations of the units involved in an unsupervised manner without extensive human involvement. We will take advantage of recent advances in *deep learning* (Bengio, 2009) to induce general representations for different modalities and learn how these interact and can be rendered in natural language. We will follow the general paradigm of *encoder-decoder* modeling (Cho et al., 2014; Sutskever et al., 2014) where an *encoder* transforms the input into a representation and a *decoder* reconstructs the input and in our case produces the corresponding translation. Advantageously, the encoder-decoder architecture can be jointly trained to maximize $P(\mathbf{x}|\mathbf{y})$, the probability of the output given the input, without having to a priori decide about the units of the translation and their correspondence. Beyond addressing a fundamental aspect of the translation problem, the proposed framework will lead to the development of novel internet-based applications that automatically simplify text, produce documentation for source code, index images with meaningful descriptions, and summarize database content in natural language.

1.1. Background

We are not aware of any previous work that addresses translation from comparable corpora in a unified framework. Existing research has tackled isolated problems using different modeling tools.

Text-to-Text Translation. We use text-to-text translation as an umbrella term for monolingual rewriting tasks that take naturally occurring texts as input and reformulate them into new texts satisfying specific constraints such as length or style. Existing approaches differ in terms of the techniques employed as well as the amount of text rewriting being performed. Most earlier work has adopted rule-based approaches (e.g., Chandrasekar et al. 1996). The success of statistical machine translation has given impetus to repurpose several SMT modeling ideas for monolingual translation tasks (e.g., Wubben et al. 2012). Unfortunately, SMT-based approaches have not proved effec-

[1] The term describes the ease with which a document can be read and understood.

tive due to the lack of large scale parallel datasets and the use of modeling assumptions that are valid for MT but often unwarranted when dealing with monolingual tasks. There has been growing interest in the development of models that are structurally aware and learn rewrite rules in terms of syntactic constituents or subtrees rather than arbitrary phrases. Such models have been applied to document simplification (Woodsend and Lapata, 2011), and summarization (Cohn and Lapata, 2013). However, their reliance on a parser limits their application to resource-rich languages, and to text-to-text translation problems.

Translation from Non-linguistic Input. The task of automatically translating between text and non-linguistic input has assumed several guises in the literature. Examples include concept-to-text generation (Reiter and Dale, 2000), image description generation (e.g., Kuznetsova et al. 2012), caption generation for complex graphical representations such as pie charts (Mittal et al., 1998) and the development of natural language interfaces to query structured information such as source code (Liu and Lieberman, 2005).

While existing concept-to-text generation systems can be engineered to obtain good performance, it is often difficult to adapt them across different domains as they rely mostly on handcrafted components. Hand engineered sentence templates are also commonly used to generate image descriptions. These are essentially syntactic patterns with slots whose fillers are words that have been identified with previously-trained visual detectors and describe objects, actions, locations, or attributes. A few models create image descriptions from scratch through a combination of phrases extracted from documents (Feng and Lapata, 2013) or human-written captions in a database (Kuznetsova et al., 2012). Rather than combining different components together each focusing on a single sub-task, we would like to develop a single joint model which can be trained on large comparable corpora consisting of images and collocated text.

Recent Advances in Deep Learning. The term *deep learning* denotes a broad family of machine learning methods focused on learning representations of data based on hierarchical artificial neural networks. The popularity of deep neural networks is due to novel training algorithms (Hinton et al., 2006) which are based on the principle of greedy layer-wise unsupervised pre-training followed by supervised fine-tuning. In NLP neural networks have seen widespread use in multiple tasks such as parsing (Collobert et al., 2011), language modeling (e.g., Mikolov et al. 2010), and sentiment analysis (Socher, 2014). An attractive aspect of deep learning methods is their ability to perform these tasks without external hand-designed resources or time-intensive feature engineering. A key concept is the notion of *embedding* which refers to the representation of symbolic information (e.g., words, sentence or documents) in terms of continuous-valued vectors. Neural networks are particularly well-suited for our translation tasks which are multimodal (feature embeddings have also proven hugely successful in large-scale visual recognition), represented by large amounts of noisy data, and without pre-existing well-engineered features. Additional reasons which have helped deep architectures obtain state of the art performance include larger datasets, parallel computers and a plethora of machine learning insights into sparsity regularization and optimization.

1.2. Research Objectives

The overall objective of this project is to develop a unified modeling framework for translating from comparable corpora. We detail below our specific objectives, each one relating to a particular Work Package (described in the Methodology section).

A. Definition of Translation Task Although multilingual translation is fairly well-understood and well-studied, this is not the case for monolingual translation and translation from non-linguistic input. Our aim is to formally characterize this translation process, to study how it manifests itself in real data, and to devise novel algorithms that gather comparable corpora according to different user requirements and task specifications.

B. Modeling Framework We propose to formalize the translation process following the *encoder-decoder* modeling paradigm. In this framework, the encoder extracts a representation of the input and the decoder reconstructs from this representation the target output. Rather than breaking up the translation problem into a sequence of local decisions (e.g., segmentation, alignment, generation), we will estimate both encoding and decoding components jointly.

C. Development of Applications We will demonstrate that our approach has practical importance by developing key applications representative of the different aspects of the translation problem. We will focus on applications that produce textual output whilst translating from linguistic or non-linguistic input.

Research Experience The PI has had extensive experience with the creation of comparable corpora from Wikipedia (Woodsend and Lapata, 2011), online news articles (Feng and Lapata, 2013), and large scale databases (Reddy et al., 2014). The first objective is a natural extension of her work that would allow a broader characterization of the translation problem which is a prerequisite to modeling. The second objective is a new and exciting direction that builds on the PI's work on developing robust models and efficient inference mechanisms for NLP tasks. Her recent work on using neural neural networks for representing and generating language can be seen as a pilot for the current proposal (Silberer and Lapata, 2014; Zhang and Lapata, 2014). As for the third objective, the PI has developed variety of text-to-text applications, including simplification (Woodsend and Lapata, 2011), summarization (Cohn and Lapata, 2013) and concept-to-text generation (Konstas and Lapata, 2013). She has also worked on problems that integrate language and vision (e.g., Feng and Lapata 2013).

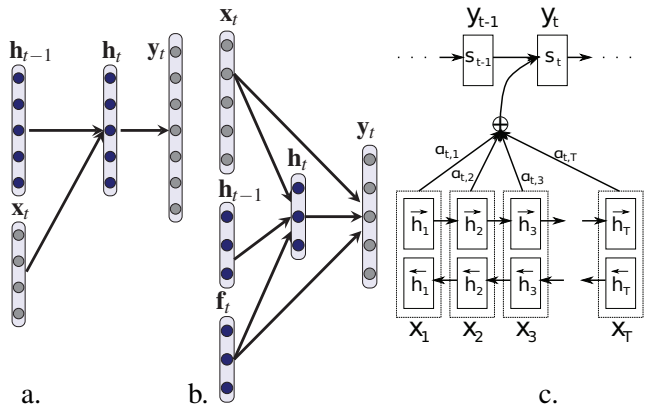

Figure 1: Examples of neural network model architectures: (a) recurrent neural network; (b) with auxiliary input layer $\mathbf{f}_t$; (c) with bidirectional RNN encoder.

2. Methodology

The proposal is organized into three Work Packages (WPs), each WP dealing with one of the objectives given in Section 1.2.

2.1. WP1: Definition of the Translation Task

This WP will formalize the translation problem by studying its different manifestations in naturally occurring data. For text-to-text translation problems, we will consider news articles and their highlights. We will also gather corpora consisting of tweets and the articles they refer to. We can create a simplification corpus by mining the Simple English Wikipedia and the main English Wikipedia. We can also explore the revision histories, to gather data pertaining to simplification, compression, and general rewriting operations. Unfortunately, there are no well-established benchmark datasets for translating from non-linguistic input. We aim to rectify this by constructing several database-to-text, source-code-to-text, and image-to-text comparable corpora. Examples include weather or sports related databases and collocated text, news articles and their images, shopping sites with product images and their descriptions, documentation for Java methods or libraries.

2.2. WP2: Modeling Framework

There are three components to our modeling framework that specify a translation model: the encoder, the decoder, and the overall network model architecture. We will first illustrate the basics of our approach and then discuss more advanced modifications and extensions.

Network Architecture Figure 1 a illustrates a typical recurrent neural network (RNN) architecture containing three layers, an input layer, a hidden layer, and an output layer (where $\mathbf{x}_t$ is the embedding of the input word at time t, $\mathbf{h}_{t-1}$ is a real-valued vector, encoding the history of all words observed in the sequence up to time step $\mathbf{t}-1$). Word embedding $\mathbf{x}_t$ is integrated with previous history $\mathbf{h}_{t-1}$ to generate the current hidden layer, which is a new history vector $\mathbf{h}_t$. Based on $\mathbf{h}_t$, we can predict the probability of the next word, which forms the output layer $\mathbf{y}_t$. The new history $\mathbf{h}_t$ is used for the future prediction, and updated with new information from word embedding $\mathbf{x}_t$ recurrently. The RNN network in Figure 1a is *not* a translation model. It defines a language model which can be used as a generator, i.e., to predict grammatically coherent word sequences, without however capturing the conditional dependence of the output text given some input. A straightforward way to model the translation task is to add add *auxiliary* input $\mathbf{f}_t$ to the recurrent neural network language model. With this extension, the RNN can measure the consistency between the source and its target in a context-sensitive way (e.g., Mikolov and Zweig 2012). The auxiliary input layer in Figure 1b can be used to feed in information appropriate for our tasks. For example, in sentence simplification $\mathbf{f}_t$ represents complex sentences, in summarization $\mathbf{f}_t$ is a document, and in image description generation $\mathbf{f}_t$ is an encoding of the information contained in the image. In the encoder-decoder framework, the encoder reads $\mathbf{f}_t$, and the decoder is often trained to predict the next word given all previously predicted words, the hidden layer configuration $\mathbf{h}_t$, and the auxiliary input layer configuration $\mathbf{f}_t$.

In Figure 1b, the auxiliary input layer $\mathbf{f}_t$ is a fixed-length vector per input sentence or more generally per input-output pair. This may turn out to be too limiting for problems with loose source-target correspondence. Content selection only implicitly takes place via $\mathbf{f}_t$ and its associations with the target layer. An architecture which implements content selection explicitly is shown in Figure 1c (Bahdanau et al., 2014). Here, the source is encoded into a sequence of vectors $x_1 \ldots x_T$, and a subset of these is chosen adaptively while decoding the input which could correspond to words, sentences, database entries, code snippets, or even regions in images. This avoids the problem of having to encode all available information into a fixed-length vector. Instead, the probability of generating output word y_t is conditioned on a sequence of hidden *annotations* $h_1 \ldots h_T$ to which the encoder maps the source input. The annotations summarize the information found in preceding *and* following words, with weights expressing which parts of the input are important in generating the target output.

The architectures in Figure 1 specify a general modeling framework; depending on how the encoder is specified and how the decoder estimates the conditional probability $p(\mathbf{x}|\mathbf{y})$ a variety of models can be expressed. It is a research question to establish the best model architecture, encoder, and decoder for each input (e.g., text versus images) We give examples of candidates we plan to develop in this project below.

Encoder In sentence-based text-to-text translation tasks (e.g., simplification, compression), $\mathbf{h}_t$ (see Figure 1) encodes information about the meaning of the source sentence, whereas $\mathbf{f}_t$ represents auxiliary information (also sentence-based). Various ways of encoding sentences have been described in the literature. A general class of basic models consists of a pro-

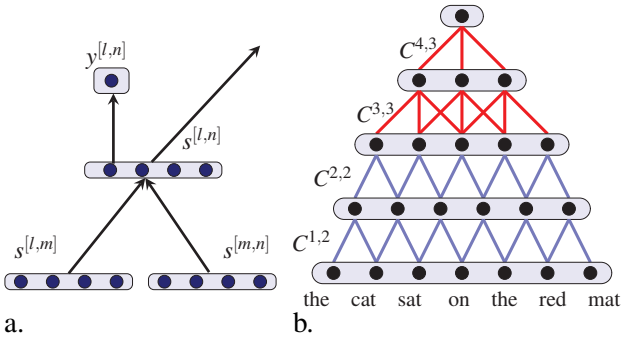

Figure 2: Examples of encoders: (a) recursive neural network; (b) convolutional sentence model (the first layer has seven vectors, one for each word. Two neighboring vectors are merged to one vector in the second layer with weight matrix $C^{1,2}$. In other layers, either two or three neighboring vectors are merged.)

jection layer that maps words, sub-word units, or n-grams to high dimensional embeddings; the latter are subsequently combined componentwise with an operation such as summation. A model that adopts a more general structure provided by an external parse tree is the Recursive Neural Network (Pollack, 1990). At every node in the tree the contexts at the left and right children of the node (denoted by $s^{[l,m]}$ and $s^{[m,n]}$ in Figure 2a) are combined by a classical layer . The weights of the layer are shared across all nodes in the tree. The layer computed at the top node gives a representation for the sentence. In Figure 2a, $s^{[l,n]}$ is the generated representation of the parent node (which could be a noun phrase or a verb phrase) as well as the representation of the entire sub-tree spanning positions l to n. $y^{[l,n]}$ is a score indicating how plausible the new node would be. Recursive neural networks have been successfully used to represent images and sentences (Socher, 2014) and could also conceivably encode documents, e.g., using the output of a text-level discourse parser (Feng and Hirst, 2014).

A third class of encoders is based on the convolution operation (Kalchbrenner and Blunsom, 2013). A convolutional sentence model (CSM) creates a representation for a sentence that is progressively built up from representations of the n-grams in the sentence. Although the CSM does not make use of an explicit parse tree, the operations that generate the representations act locally on small n-grams in the lower layers of the model and act increasingly more globally on the whole sentence in the upper layers of the model. A graphical illustration of a CSM is shown in Figure 2b. The model learns an array of weight matrices which compress neighboring columns in the current layer to one column in the next layer. The CSM embodies hierarchical structure, without relying on parse trees. It can be therefore robustly applied across languages, genres, and modalities. Indeed, deep Convolutional neural networks are nowadays the architecture of choice for large-scale image recognition (Simonyan and Zisserman, 2014) and have been used to represent the meaning of documents (Denil et al., 2014).

Decoder In the architectures sketched above, the decoder is trained to predict the next word given a context vector and all previously predicted words (Figure 1b) or alternatively given a *distinct* context vector for each target word (Figure 1c). In both instances, the decoder is implemented using an RNN. This choice has been successfully used to model bilingual machine translation (e.g., Kalchbrenner and Blunsom 2013) and image description generation (Vinyals et al., 2014). We propose to experiment with simpler decoders, where the model is asked to learn how to associate a given input to only parts of the output, e.g., phrase representations such as verb phrases, noun phrases, and so on. The advantages are computational (since such models will be easier to train) but also theoretical since they allow to explore different ways of generating text. For instance the phrases could be arranged into text using templates or more sophisticated methods based on graphs, dependencies or even integer linear programming. Importantly, this would allow injecting task specific constraints into the generation process pertaining to length or style. We will also experiment with more sophisticated decoders, which take syntactic information into account either in the form of a linearized parse tree, or head-modifier dependencies.

2.3. WP3: Development of Applications

We will use several applications as a means of evaluating our framework and showcasing its practical utility. We will develop three text-to-text applications, namely simplification, generation of story highlights, and tweet messages for news articles. We will also deploy systems that translate from non-linguistic input focusing on: (a) the automatic generation of documentation for source code; (b) translation from databases into natural language for the weather, sports, and accidents (e.g., earthquakes) domains; and (c) image description generation for generic images and domain specific ones (e.g., found in consumer sites). All the tasks discussed in this proposal will be subject to the same evaluation protocol. We will employ standard automatic evaluation measures (e.g., ROUGE, BLEU) during system development, whereas stable system versions will be assessed with real users.

2.4. Expected Impact

The practicality and range of many NLP applications would be significantly enhanced if one could automatically reformulate linguistic and non-linguistic content for a variety of domains and text genres. The algorithmic and modeling developments of this project will be of interest to the NLP, Machine Learning, Computer Vision, as well as software engineering communities. Moreover, being able to translate images and other forms of non-linguistic information into text, will render this type of data more amenable to search engines, and more generally to any type of software that expects text as input (e.g., speech synthesizers, screen readers).

Overall, the proposed research will render the internet more accessible to a broader audience by: (a) facilitating the development of reading aids for a wide range of users (e.g., low-literacy readers, non-native speakers), (b) reducing the information overload that has resulted from the proliferation of online data and (c) alleviating the problem of accessing non-textual information on the web for sighted and visually impaired users.

3. Resources

The proposed project requires a methodologically diverse team, with expertise in NLP, machine learning, algorithms and data structures, and image processing. The PI will devote 70% of her time to the project. She will take an active (not merely supervisory) role in the delivery of all work packages which build on her previous research. In addition, she will provide scientific leadership and management throughout the project.

The research team will include two postdoctoral researchers. One postdoc with NLP background will be in charge of the text-to-text translation side. The second postdoc will have expertise in machine learning and computer vision and will be responsible for developing translation models from non-linguistic data. The postdocs will overlap by two years so as to maximize collaboration and the goal of developing a unified translation framework. Three PhD students are also part of the team. One of them will focus on the image description generation application and the other one on the automatic generation of documentation for source code. The two applications address fundamental research questions (e.g., what is the relationship between other modalities and language, how can we translate between them) and are thus ideally suited for individual PhD theses. The third student will work on representation learning for translation tasks which is an important part of the proposed work.

We request 15% of a computing officer to deal with the complex computing requirements posed by the modeling work. An administrator (10%) will provide dedicated secretarial and administrative support for the management of this project. Resources for paying annotators and experimental participants are also requested and a travel budget to cover attendance to conferences and research visits. Finally, the proposed work is highly data-intensive, and thus funds are requested for purchasing GPUs and builiding a robust computing infrastructure for experimentation.

References

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. arXiv:1409.0473.

Bengio, Y. (2009). Learning deep architectures for AI. foundations and trends in machine learning. *Foundations and Trends in Machine Learning*, 2(1):1–127.

Chandrasekar, R., Doran, C., and Srinivas, B. (1996). Motivations and methods for text simplification. In *Proceedings of the 16th COLING*, pages 1041–1044, Copenhagen, Danemark.

Cho, K., van Merrienboer, B., Bahdanau, D., and Bengio, Y. (2014). On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8*, pages 103–111, Doha, Qatar.

Cohn, T. and Lapata, M. (2013). An abstractive approach to sentence compression. *ACM Transactions on Intelligent Systems and Technology*, 4(3):1–35.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12:2493–2537.

Denil, M., Demiraj, A., and de Freitas, N. (2014). Extraction of salient sentences from labelled documents. arXiv:1412.6815.

Feng, V. W. and Hirst, G. (2014). A linear-time bottom-up discourse parser with constraints and post-editing. In *Proceedings of the 52nd ACL*, pages 511–521, Baltimore, Maryland.

Feng, Y. and Lapata, M. (2013). Automatic caption generation for news images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(4):797–812.

Hinton, G. E., Osindero, S., and Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, 18(1527–1554).

Kalchbrenner, N. and Blunsom, P. (2013). Recurrent continuous translation models. In *Proceedings of the 2013 EMNLP*, pages 1700–1709, Seattle, Washington, USA.

Konstas, I. and Lapata, M. (2013). A global model for concept-to-text generation. *Journal of Artificial Intelligence Research*, 48:305–346.

Kuznetsova, P., Ordonez, V., Berg, A., Berg, T., and Choi, Y. (2012). Collective generation of natural image descriptions. In *Proceedings of the 50th ACL*, pages 359–368, Jeju Island, Korea.

Liu, H. and Lieberman, H. (2005). Metafor: visualizing stories as code. In *Proceedings of the 2005 International Conference on Intelligent User Interfaces*, pages 305–307, San Diego, CA.

Mikolov, T., Karafiát, M., Burget, L., Cernockỳ, J., and Khudanpur, S. (2010). Recurrent neural network based language model. In *INTERSPEECH*, pages 1045–1048.

Mikolov, T. and Zweig, G. (2012). Context dependent recurrent neural network language model. In *SLT*, pages 234–239.

Mittal, V. O., Moore, J. D., Carenini, G., and Roth, S. (1998). Describing complex charts in natural language: A caption generation system. *Computational Linguistics*, 24:431–468.

Oard, D. W. and Och, F. J. (2003). Rapid-response machine tranlsation for unexpected languages. In *Proceedings of MT Summit IX*, New Orleans, LA.

Pollack, J. B. (1990). Recursive distributed representations. *Artificial Intelligence*, (46):77–105.

Reddy, S., Lapata, M., and Steedman, M. (2014). Large-scale semantic parsing without question-answer pairs. *Transactions of the ACL*. To appear.

Reiter, E. and Dale, R. (2000). *Building natural language generation systems*. Cambridge University Press, New York, NY.

Silberer, C. and Lapata, M. (2014). Learning grounded meaning representations with autoencoders. In *Proceedings of the 52nd ACL*, pages 721–732, Baltimore, Maryland.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recogntion. arXiv:1409.1556.

Socher, R. (2014). *Recursive Deep Learning for Natural Language Processing and Comptuer Vision*. PhD thesis, Stanford University.

Sutskever, I., Vinyals, O., and Le, Q. (2014). Sequence to sequence learning with neural networks. arXiv:1409.3215.

Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2014). Show and tell: A neural image caption generator. arXiv:1411.4555.

Woodsend, K. and Lapata, M. (2011). Learning to simplify sentences with quasi-synchronous grammar and integer programming. In *Proceedings of the 2011 EMNLP*, pages 409–420, Edinburgh, Scotland, UK.

Wubben, S., van den Bosch, A., and Krahmer, E. (2012). Sentence simplification by monolingual machine translation. In *Proceedings of the 50th ACL)*, pages 1015–1024, Jeju Island, Korea.

Zhang, X. and Lapata, M. (2014). Chinese poetry generation with recurrent neural networks. In *Proceedings of the 2014 EMNLP*, pages 670–680, Doha, Qatar.

b. Curriculum Vitae

Education

2001 **PhD in Informatics, University of Edinburgh**. Title of dissertation: *The Acquisition and Modelling of Lexical Knowledge: A Corpus-based Investigation of Systematic Polysemy*. Supervisors: Alex Lascarides, Chris Brew, Steve Finch.

1997 **Master in Language Technologies, Carnegie Mellon University**. Coursework: parsing, generation, algorithms, statistical NLP, machine translation, computational models of neural systems, machine learning, software engineering, speech recognition/synthesis, theoretical linguistics, formal semantics, cognitive psychology.

1995 **Master in Cognitive Science and Natural Language, University of Edinburgh**. Coursework: cognitive psychology, computational linguistics, syntax, formal semantics, knowledge representation, formal logic, logic programming, neural computation.

1994 **Bachelor in Linguistics, Athens University**. Coursework: syntax, semantics, morphology, pragmatics, psycholinguistics, sociolinguistics, speech recognition, computational linguistics, formal logic.

Employment

2013 **Visiting Professor, MIT Computer Science & Artificial Intelligence Lab**. Hosted by Prof. Regina Barzilay; collaborations with MIT and Harvard faculty on models of grounded language acquisition and representation.

2012 **Professor, School of Informatics, University of Edinburgh**. Working on text-to-text generation, meaning representation, and interface of language and vision using unsupervised learning paradigms.

2007 **Reader, School of Informatics, University of Edinburgh**. Established research group in natural language processing and generation; diversified into grounded models of language

2005 **Lecturer, School of Informatics, University of Edinburgh**. Diversified research into unsupervised learning, with post-doctoral researchers and PhD students.

2003 **Lecturer, Department of Computer Science, University of Sheffield**. Built research portfolio in natural language processing.

2001 **Research Fellow, Department of Computational Linguistics, Saarland University**. Postdoctoral work on statistical language processing and natural language generation; acquired skills in probabilistic modeling and experimental paradigms.

Fellowships and Awards

2010 Karen Spärck Jones Award, British Computer Society and Information Retrieval Specialist Group.

2007 Best paper award, Conference on Empirical Methods in Natural Language Processing and on Computational Natural Language Learning, Prague.

2005 Statistical Models for Text-to-text Generation, EPSRC Advanced Fellowship (5 years). £457,914.

2002 Best paper award, Conference on Empirical Methods in Natural Language Processing, Philadelphia.

Research and Postgraduate Supervision

Research Fellows	Oier Lopez de Lacalle (2012–2013), Kristian Woodsend (2009–now), Trevor Cohn (2006–2009), Caroline Sporleder (2003–2005).
Current PhD students	Philip Gorinski (2013–now), Xingxing Zhang (2013–now), Lea Frermann (2013–now), Siva Reddy (2012–now), Carina Silberer (2011–now), William Blacoe (2011–now).
PhD students graduated	Ioannis Konstas (2009–2013; now postdoctoral researcher, School of Informatics, University of Edinburgh), Trevor Fountain (2008–2013, now co-founder Blazing Griffin), Joel Lang (2008–2012, now postdoctoral researcher, Department of Computer Science, University of Geneva), Yansong Feng (2006–2010, now lecturer, Peking University), Neil McIntyre (2006–2010, now language engineer, Analog Devices, Edinburgh), Jeffrey Mitchell (2006–2010; now postdoctoral researcher, School of Informatics, University of Edinburgh), Hagen Fürstenau (2008–2010, now postdoctoral researcher, Columbia University), Samuel Brody (2005–2008, now software engineer, Google), James Clarke (2004–2007, now manager, Basis Technology), Sebastian Padó (2002–2007, now professor, University of Heidelberg).

Professional Service

Associate editor	Transactions of the Association for Computational Linguistics (2012–2014), Transactions on Speech and Language Processing (2008–2013), Journal of Artificial Intelligence Research (2010–now).
Editorial board	Journal of Artificial Intelligence Research (2006–2008), Computational Linguistics (2002–2004).
Program Chair	Conference of the European chapter of the Association for Computational Linguistics (2012), Conference on Computational Natural Language Learning (2010), Conference on Empirical Methods on Natural Language Processing (2008).
Area chair	Annual Meeting of the Association for Computational Linguistics and International Joint Conference of the Asian Federation of Natural Language Processing (2015), Conference on Empirical Methods on Natural Language Processing (2011), International Joint Conference on Artificial Intelligence (2011), International Joint Conference on Artificial Intelligence (2009), International Conference on Machine Learning (2009), International Joint Conference on Artificial Intelligence (2009), Conference of the European chapter of the Association for Computational Linguistics (2006), Association for Computational Linguistics (2005).
Journal reviewer	Computational Linguistics, Cognitive Science, Journal of Artificial Intelligence Research, ACM Transactions on Speech and Language Processing, Computational Intelligence, Journal of Machine Learning Research, Computer Speech and Language.
Conference reviewer	Association for Computational Linguistics (2001–now), Empirical Methods on Natural Language Processing (2001–now), European chapter of the Association for Computational Linguistics (1999–now), North American Chapter of the Association for Computational Linguistics (2001–now).
Grant reviewer	Engineering and Physical Sciences Research Council (EPSRC), Economic and Social Research Council (ESRC), US National Science Foundation (US NSF).

Appointments as External Examiner

2015	Nikhil Garg, PhD *Unsupervised Semantic Role Labeling*, University of Geneva.
2014	Jackie Chi Kit Cheung, PhD *Towards Large-Scale Natural Language Inference with Distributional Semantics*, University of Toronto.
2014	Alona Fyshe, PhD, *Decoding Word Semantics from Magnetoencephalography Time Series Transformations*, Carnegie Mellon University.
2012	Justin Washtell, PhD, *Towards a Purely Distributional Model of Meaning: Distance, Expectation, and Composition*, University of Sussex.
2011	Hagen Fürstenau, *Semi-supervised Semantic Role Labeling via Graph Alignment*, PhD, Saarland University.
2009	Katja Filippova, *Dependency Graph Based Sentence Fusion and Compression*, PhD, University of Heidelberg.
2008	Yee Seng Chan, *Word Sense Disambiguation: Scaling up, Domain adaptation, and Application to Machine Translation*, PhD, National University of Singapore.
2007	Olga Uryupina, *Knowledge Acquisition for Coreference Resolution*, PhD, Saarland University.
2005	Judita Preiss, *Probabilistic Word Sense Disambiguation*, PhD, University of Cambridge.
2003	Julie Weeds, *Measures and Applications of Lexical Distributional Similarity*, PhD, University of Sussex.

Career Breaks

2006–2007	Maternity leave
2010–2011	Maternity leave

Appendix: All on-going and submitted grants and funding of the PI (Funding ID)

On-going Grants

Project Title	Funding Source	Amount (Euros)	Period	Role of the PI	Relation to current ERC proposal
Readers: Evaluation and Development of Reading Systems	EPSRC (CHIST-ERA)	€1,193,934	2013–2016	PI	developed neural network models for learning representations which may serve as encoders in modeling framework
A Unified Model of Compositional and Distributional Semantics: Theory and Applications	EPSRC	€2,043,861	2012–2015	PI	none
An Integrated Model of Syntactic and Semantic Prediction in Human Language Processing	EPSRC	€454,795	2011–2015	co-PI	none
Statistical Models for Text-to-text Generation	EPSRC Fellowship	€399,567	2005–2010	PI	none
Robust Pragmatics for Narrative Text	EPSRC	€322,244	2002–2005	co-PI	none

Grant Applications

Project Title	Funding Source	Amount (Euros)	Period	Role of the PI	Relation to current ERC proposal
ActionNet: Building and Exploiting an Images Database Based on the WordNet Verb Hierarchy	EPSRC	€828,000	under review	co-PI	development of large-scale image database which can be used in image description generation tasks mentioned in proposal

Past Grants

Project Title	Funding Source	Amount (Euros)	Period	Role of the PI	Relation to current ERC proposal
Global Inference for Summarization Using Integer Linear Programming	EPSRC	€377,549	2008–2011	PI	developed applications which will be used as baselines for comparison with models proposed in ERC grant
Ranking Word Senses for Disambiguation: Models and Applications	EPSRC	€429,768	2008–2011	PI	none
Application-based Text-to-Text Generation	EPSRC	€232,355	2006–2009	PI	none

c. Early Achievements Track-record

Mirella Lapata is a professor at the School of Informatics at the University of Edinburgh. She holds a personal chair in Natural Language Processing. Her research focuses on machine learning for natural language understanding and generation. At the most general level her contribution has been in developing computational models for the representation, extraction, and generation of semantic information from structured (e.g., databases) and unstructured data (e.g., web-scale corpora). She has worked on a variety of applied NLP tasks such as word sense disambiguation, distributional models of word meaning, semantic parsing and semantic role labeling, discourse coherence, summarization, text simplification, concept-to-text generation, and question answering. She has also used computational models (drawing mainly on probabilistic generative models) to explore aspects of human cognition such as learning concepts, judging similarity, forming perceptual representations, and learning word meanings. A brief summary of her research contributions is given below with an emphasis on areas specifically related to the current proposal.

Unsupervised Learning Lapata has introduced novel learning algorithms for NLP tasks that traditionally relied on hand-coded knowledge. Her work has been instrumental in opening up areas of NLP for experimental and data-driven treatment which so far resisted such methods, in particular in natural language generation and computational semantics. For example, her group pioneered the use of the web as a large corpus and demonstrated the overwhelmingly better performance of web-scale models for a wide range of NLP tasks. This work was featured in the Economist, and received the Best Paper Award at EMNLP 2002. It has had a sustained impact and helped to create a new research field Web as Corpus, which now has its own annual workshop (WAC), its own special interest group (ACL SIGWAC), and a separate session at major conferences (e.g., the AI and the Web track at AAAI). Another example is her work on sentence compression (i.e., the task of creating a shorter version of a sentence by removing extraneous information). Previous solutions to the compression problem have been cast mostly in a supervised learning setting. Lapata showed that a formulation of sentence compression as an optimization problem requires minimal supervision, is robust across domains, and extends to document compression. This research won the Best Paper Award at EMNLP 2007, and has helped to establish Integer Linear Programming as a standard machine learning paradigm for NLP with is own specialized workshop.[2]

Learning Representations A fundamental question is NLP is how to represent linguistic and domain knowledge in order to facilitate learning. As an example, concept-to-text generation is an application which has traditionally relied on extensive manual effort. Lapata's work shows that some of the components involved can be *learned* by exploiting the vast resource of documents available on the web or by reconceptualizing the problem in a way that renders it more amenable to learning. Discourse coherence is a case in point, where systems often rely on elaborate rules in order to ensure that the generated documents are understandable. It is however possible to represent a document as an *entity grid*, a two-dimensional array that captures the distribution of discourse entities across text sentences. This novel entity-based representation of discourse allows to learn the properties of coherent texts from a corpus, without any hand-coding. Another example is her work on vector-based models of semantic composition. She has developed one of the first empirical frameworks of composition where word meaning is represented quantitatively with composition operations similar to those found in logic-based accounts of semantics. This research has received significant attention both from cognitive scientists and NLP practitioners. In the past years, two ACL workshops were devoted to this topic, Disco-2011 and GEMS-2010 both of which organized shared tasks and used data that originated in her work.

Language Grounding More recently, she has developed an interest in grounded language acquisition, where the goal is to extract representations of the meaning of natural language tied to the physical world. She has specifically worked on two instantiations of the grounding problem, i.e., associating language to perceptual data such as images and mapping natural language expressions (e.g., questions) to machine interpretable formal meaning representations. Her research group has published the first two papers on multimodal semantics and image description generation. For her contributions to the emerging area of vision and language, she has been invited to present her work at NIPS (2011), ACL (2013), CVPR (2014) and to the University of Washington and Microsoft Research Summer Institute (2013).

Prof. Lapata has maintained a high publication output including 30 peer-reviewed journal articles and 74 peer-reviewed papers in conference proceedings. These are high-impact journals (e.g., *Computational Linguistics*, *Journal of Artificial Intelligence Research*) and highly competitive conferences with an acceptance rate as low as 20% (e.g., ACL, AAAI). Her total citation count is 6,017 and her h-index is 45 (based on Google Scholar). In addition to publishing peer-reviewed work, she maintains a busy schedule of invited lectures at conferences, workshops and research centres around the world. She has given six keynote talks in the past three years, and served as program co-chair for three major NLP conferences (EACL, CoNLL, and EMNLP). To date, she has successfully graduated ten PhD students all of whom have leading positions in industry or academia, and is currently supervising PhDs and one postdoctoral fellow. Over the past 10 years, Lapata has managed a portfolio of more than €2.4 million as a PI and Co-I.

[2] http://ilpnlp.wikidot.com/naacl-hlt-workshop

Selected Journal Publications[3]

Reddy, Siva, Mirella Lapata, and Mark Steedman. 2014. Large-scale Semantic Parsing without Question-Answer Pairs. *Transactions of the ACL* To appear.

Lang, Joel, and Mirella Lapata. 2014. Similarity-Driven Semantic Role Induction via Graph Partitioning. *Computational Linguistics* 40(3): 133–164. (**62**).

Feng, Yansong, and Mirella Lapata. 2013. Automatic Caption Generation for News Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35(4): 797–812. (**55**).

Konstas, Ioannis, and Mirella Lapata. 2013. A Global Model for Concept-to-Text Generation. *Journal of Artificial Intelligence Research* 48: 305–346. (**39**).

Mitchell, Jeff, and Mirella Lapata. 2010. Composition in Distributional Models of Semantics. *Cognitive Science* 34(8): 1388–1429. (**208**).

Navigli, Roberto, and Mirella Lapata. 2010. An Experimental Study of Graph Connectivity for Unsupervised Word Sense Disambiguation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32(4): 678–692. (**146**).

Padó, Sebastian, and Mirella Lapata. 2009. Cross-lingual Annotation Projection for Semantic Roles. *Journal of Artificial Intelligence Research* 36: 307–340. (**55**).

Cohn, Trevor, and Mirella Lapata. 2009. Sentence Compression as Tree Transduction. *Journal of Artificial Intelligence Research* 34: 637–674. (**90**).

Clarke, James, and Mirella Lapata. 2008. Global Inference for Sentence Compression: An Integer Linear Programming Approach. *Journal of Artificial Intelligence Research* 31: 273–381. (**140**).

Barzilay, Regina, and Mirella Lapata. 2008. Modeling Local Coherence: An Entity-based Approach. *Computational Linguistics* 34: 1–34. (**316**).

Padó, Sebastian, and Mirella Lapata. 2007. Dependency-based Construction of Semantic Space Models. *Computational Linguistics* 33(2): 161–199. (**422**).

Lapata, Mirella. 2006. Automatic Evaluation of Information Ordering. *Computational Linguistics* 32(4): 471–484. (**106**).

Lapata, Mirella, and Frank Keller. 2005. Web-based Models for Natural Language Processing. *ACM Transactions on Speech and Language Processing* 2: 1–31. (**187**).

Lapata, Mirella, and Chris Brew. 2004. Verb Class Disambiguation Using Informative Priors. *Computational Linguistics* 30(1): 45–73. (**69**).

Lapata, Mirella, and Alex Lascarides. 2006. Learning Sentence-internal Temporal Relations. *Journal of Artificial Intelligence Research* 27: 85–117. (**68**).

Keller, Frank, and Maria Lapata. 2003. Using the Web to Obtain Frequencies for Unseen Bigrams. *Computational Linguistics* 29(3): 459–484. (**365**).

Selected Conference Publications

Gorinski, Philip, and Mirella Lapata. 2015. Movie Script Summarization as Graph-based Scene Extraction. In *Proceedings of NAACL*. Denver, Colorado. To appear.

Ortiz, Luis Gilberto Mateo, Clemens Wolff, and Mirella Lapata. 2015. Learning to Interpret and Describe Abstract Scenes. In *Proceedings of NAACL*. Denver, Colorado. To appear.

Zhang, Xingxing, and Mirella Lapata. 2014. Chinese Poetry Generation with Recurrent Neural Networks. In *Proceedings of the 2014 EMNLP*, 670–680. Doha, Qatar.

Silberer, Carina, Vittorio Ferrari, and Mirella Lapata. 2013. Models of Semantic Representation with Visual Attributes. In *Proceedings of the 51st ACL*, 572–582. Sofia, Bulgaria. (**17**).

Blacoe, William, and Mirella Lapata. 2012. A Comparison of Vector-based Representations for Semantic Composition. In *Proceedings of the 2012 EMNLP and CoNLL*, 546–556. Jeju Island, Korea: Association for Computational Linguistics. (**51**).

Woodsend, Kristian, and Mirella Lapata. 2011. Learning to Simplify Sentences with Quasi-Synchronous Grammar and Integer Programming. In *Proceedings of the 2011 EMNLP*, 409–420. Edinburgh. (**60**).

Feng, Yansong, and Mirella Lapata. 2010. How Many Words Is a Picture Worth? Automatic Caption Generation for News Images. In *Proceedings of the 48th ACL*, 1239–1249. Uppsala, Sweden. (**43**).

Brody, Samuel, and Mirella Lapata. 2009. Bayesian Word Sense Induction. In *Proceedings of the 12th EACL*, 103–111. Athens, Greece. (**102**).

Mitchell, Jeff, and Mirella Lapata. 2008. Vector-based Models of Semantic Composition. In *Proceedings of ACL-08: HLT*, 236–244. Columbus, Ohio. (**269**).

Shen, Dan, and Mirella Lapata. 2007. Using Semantic Roles to Improve Question Answering. In *Proceedings of the 2007 EMNLP-CoNLL*, 12–21. (**179**).

Cohn, Trevor, and Mirella Lapata. 2007. Machine Translation by Triangulation: Making Effective Use of Multi-Parallel Corpora. In *Proceedings of the 45th ACL*, 728–735. Prague. (**72**).

Lapata, Mirella, and Regina Barzilay. 2005. Automatic Evaluation of Text Coherence: Models and Representations. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, 1085–1090. Edinburgh. (**98**).

Barzilay, Regina, and Mirella Lapata. 2005. Collective Content Selection for Concept-to-Text Generation. In *Proceedings of HLT-EMNLP*, 331–338. Vancouver, British Columbia, Canada. (**39**).

Lapata, Mirella. 2003. Probabilistic Text Structuring: Experiments with Sentence Ordering. In *Proceedings of the 41st ACL*, 545–552. Sapporo, Japan. (**220**).

[3]Number of citations are based on Google Scholar and shown in bold face.