ERC Consolidator Grant 2015
Research Proposal [Part B2]

Part B2: ***The Scientific Proposal***

a. State of the Art and Objectives

Natural Language Processing (NLP), also known as Computational Linguistics, 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 an approach to machine translation (MT) that 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 a very short period of time — as little as a day (Oard and Och, 2003). In less than two decades, SMT has come to dominate academic MT research, and has gained a share of the commercial MT market. Publicly available systems like Google Translate can instantly translate between any pair of over fifty human languages and thus allow users to read web content that would otherwise 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 the most important highlights from documents or create textual descriptions for non-linguistic data such as databases, graphs, 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 automatically learning how to translate into natural language 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 objective 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 are 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 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,

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

Source	Target
Marie was born to King Stephen of England and his wife Matilda I, Countess of Boulogne.	Marie was the daughter of Stephen of England and Matilda of Boulogne.
At an early age, she was apparently placed in a convent, but she became her childless brother William's heir in 1159.	She was placed in a convent when she was young so that she could become a nun. When her brother, William of Blois, died, she became the heir of the Count of Boulogne.
Since she was the heiress to the county of Boulogne, she was forced to leave her convent and married off to Matthew of Alsace (c.1130–1173), who would become Count of Boulogne and co-ruler (1160) through his marriage to her.	She had to leave the convent and was married to Matthew of Alsace.

Table 1: Examples of source sentences and their simplifications, taken from the article *Marie I, Countess of Boulogne* in Main English and Simple English Wikipedia.

Source	Target
The first new track from David Bowie in a decade is not eligible for the UK singles chart.	David Bowie's new single "not eligible" for UK chart.
A training aircraft has crash landed near RAF Cranwell in Lincolnshire, the Ministry of Defence has confirmed. The Tutor aircraft made a forced landing in a field after a routine training sortie at about 12.30 GMT. Both crew members are said to have walked away with no serious injuries.	Plane crash lands at RAF Cranwell, Lincolnshire — no serious injuries reported.

Table 2: Examples of source sentences and their corresponding tweets, taken from the BBC news web site.

the proposed research 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. 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 (such as sentence simplification or the translation of database entries into natural language) using different approaches and modeling tools. It is outside the scope of this proposal to give a detailed overview of SMT which remains an area of much active research. We refer the interested reader to Koehn (2009) for a detailed treatment and highlight below aspects of the SMT problem that relate to our proposal.

Statistical Machine Translation. SMT was originally formulated as a series of probabilistic models that learn word-to-word correspondences from sentence-aligned bilingual parallel corpora. Current methods have focused on modeling translation based on contiguous strings of words, called *phrases* in the source language and corresponding phrases in the target language (Koehn et al., 2003). The phrase-based approach has proved to be very successful, and many state-of-the-art machine translation systems are based on it. Phrase-based models typically start by word-aligning a bilingual parallel corpus and then proceed to extract multi-word phrases (that are consistent with the alignments) which they use to build new translations. A variety of parameters are estimated using the parallel corpora including various heuristics for extracting the phrase pairs, feature functions associated with phrase translation probabilities, reordering probabilities, *n*-gram language model scores, and so on. The features are usually combined in a log linear model and their weights are set through minimum error rate training (Och, 2003).

A new class of translation models based purely on neural networks has been recently proposed (Cho et al., 2014; Kalchbrenner and Blunsom, 2013). Unlike traditional phrase-based translation systems which consist of many small sub-components that are tuned separately, neural machine translation attempts to build and train a single, large neural network that reads a sentence and outputs a correct translation. The emergence of this type of translation models is highly significant, both practically and theoretically. These models provide substantial flexibility in conceptualizing the translation problem and require only a fraction of the memory needed by traditional SMT models.

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. Examples include modeling paraphrase relationships between sentences or phrases (Barzilay, 2003), simplifying text by identifying utterances in a document that pose reading difficulty and substituting them with simpler alternatives (Chandrasekar et al., 1996), and rendering sentences shorter with minimal information loss while preserving their grammaticality (Jing, 2000). Tables 1 and 2 show naturally occurring examples of simplification and compression. Approaches differ with regard to the techniques employed as well as the amount of text rewriting performed. This is partly due to the fact that different rewriting applications impose slightly different modeling requirements but also due to the lack of readily available corpora for modeling purposes.

Most earlier work on text-to-text translation has adopted rule-based approaches (Carroll et al., 1999; Chandrasekar et al., 1996; Jing, 2000; Siddharthan, 2006). The success of SMT has given impetus to repurpose several SMT modeling ideas for monolingual translation tasks. Unfortunately, out-of-the-box SMT techniques often yield sub-optimal models that make many simplifying assumptions (e.g., to only perform monotone translation without reordering) or introduce ad-hoc post-processing steps that manipulate the model's output rather the model itself. As a result, the vast majority of existing work has focused on specialized solutions addressing specific cases of the translation problem. For instance, sentence compression is often modeled as word or constituent deletion (Knight and Marcu, 2002) whereas sentence simplification is approximated by syntactic transformations such as sentence splitting (Bach et al., 2011) or lexical substitutions (Wubben et al., 2012) but not both.

More recently, 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 are expressive enough to model any type of rewrite operation, and have been applied to a few monolingual translation tasks, including simplification (Woodsend and Lapata, 2011; Zhu et al., 2010), compression (Cohn and Lapata, 2013; Ganitkevitch et al., 2011), and summarization (Woodsend and Lapata, 2012). However, their reliance on a parser limits their application to resource-rich languages and to text-to-text translation problems. In this proposal we wish to recast the rewriting task in a deep learning framework and investigate whether structural information (when available) is beneficial (e.g., in capturing stylistic conventions and regularities).

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 (Kulkarni et al., 2011), caption generation for complex graphical representations such as pie charts (Mittal et al., 1998) and the development of natural language interfaces to query source code (Liu and Lieberman, 2005).

A typical concept-to-text generation system implements a pipeline architecture consisting of three core stages, namely text planning (determining the content and structure of the target text), sentence planning (determining the structure and lexical content of individual sentences), and surface realization (rendering the specification chosen by the sentence planner into a surface string). Traditionally, these components are hand-engineered in order to generate high quality text, however at the expense of portability and scalability. Recent years have witnessed a growing interest in automatic methods for creating trainable generation components (Barzilay and Lapata, 2005; Duboue and McKeown, 2002) or even systems that implement concept-to-text generation end-to-end (Angeli et al., 2010; Konstas and Lapata, 2012a,b). Such systems are trained on comparable corpora of databases and collocated text. Table 3 gives an example.

There has been a recent surge of interest in the development of models that automatically describe image content in natural language. Several methods leverage recent advances in computer vision and generate novel sentences relying on object detectors, attribute predictors, action detectors, and pose estimators. Generation is performed using templates or syntactic rules which piece the description together while exploiting word-co-occurrence statistics (Mitchell et al., 2012). A large body of work has focused on the complementary problem of matching sentences or phrases to an image from existing human authored descriptions (Hodosh et al., 2013). Much like previous SMT work, most approaches attempt to combine different components together each focusing on a single sub-task. In contrast, we would like to develop a single *joint* model that creates image descriptions from scratch, is based on a neural network architecture, and can be trained on large comparable corpora consist ing of images and collocated text (see Table 4 for an example of a Wikipedia image and its captions).

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 (see for example Bengio 2009). The popularity of deep neural networks is due to novel training algorithms introduced by Hinton et al. (2006) which are based on the principle of greedy layer-wise unsupervised pre-training followed by supervised fine-tuning. Many variants of deep neural networks have been applied to various tasks with impressive improvements over conventional approaches. In NLP neural networks have seen application in part-of-speech tagging, named entity recognition (Collobert et al., 2011), language modeling (Bengio et al., 2003; Mikolov et al., 2010), senti-

Source:

Fumbles

PLAYER	FUM	LOST	REC	YDS
Coles	1	1	0	0
Portis	1	1	0	0
Davis	0	0	1	0
Little	0	0	1	0
Suggs	1	0	1	0
…	…	…	…	…
…	…	…	…	…

Passing

PLAYER	CP/AT	YDS	AVG	TD	INT
Brunell	17/38	192	6.0	0	0
Garcia	14/21	195	9.3	1	0
…	…	…	…	…	…

Rushing

PLAYER	REC	YDS	AVG	LG	TD
Suggs	22	82	3.7	25	1
…	…	…	…	…	…

Target: Suggs rushed for 82 yards and scored a touchdown in the fourth quarter, leading the Browns to a 17-13 win over the Washington Redskins on Sunday. Jeff Garcia went 14-of-21 for 195 yards and a TD for the Browns, who didn't secure the win until Coles fumbled with 2:08 left.

Table 3: Example of database records and the corresponding game summary from the official site of the American National Football League (NFL).

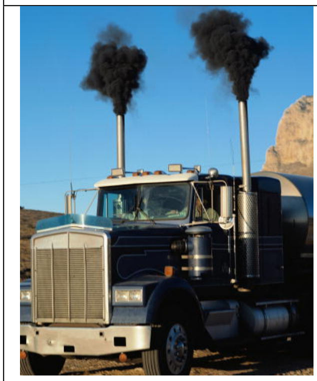
Source	Target
	Diesel trucks are the largest emitter of toxic diesel particulate matter in California. A diesel-powered truck emitting sooty exhaust gas while starting its engine. Petroleum diesel exhaust from a truck. A diesel-powered truck emits an exhaust gas rich in black particulate matter when starting its engine. Air pollution from motor vehicles is an example of a negative externally. The costs of the air pollution for the rest of society is not compensated for by either the producers or users of motorized transport. Emission of soot from a large diesel truck, without particle filters.

Table 4: Examples of images and their descriptions taken from Wikipedia.

ment analysis (Socher, 2014), and machine translation (Cho et al., 2014; Kalchbrenner and Blunsom, 2013).

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 also well-suited for tasks involving multiple modalities, (feature embeddings have also proven hugely successful in large-scale visual recognition), where well-engineered features are not available, and training data is noisy. 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.

2. Research Objectives

The overall objective of this project is to develop a unified modeling framework for translating from comparable corpora. 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) will estimate both encoding and decoding components jointly. We will adopt a broad-coverage approach and use our models to improve upon the state of the art for a wide range of NLP tasks. We detail below our specific objectives, each one relating to a particular Work Package (described in the Methodology section).

A. Definition of the 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 will develop a common modeling framework for translating from different types of input data, including text, structured databases, and images. This entails devising neural network architectures suitable for a variety of translation tasks. Specifically, we would like to develop a joint model that takes some input $\mathbf{x}$ and is trained to maximize the likelihood $p(\mathbf{x}|\mathbf{y})$, where $\mathbf{x}$ may correspond to words, images, a database, or source code and $\mathbf{y}$ is its rendering in natural language.

In order to establish what kind of representations are suitable for different translation tasks, we will perform a systematic exploration using increasingly sophisticated levels of structural information (see the Methodology section for more details). We will devise encoders which extract representations from individual sentences, documents, images, or databases using modality specific information. For instance, the representation for sentences could be based on words or n-grams but also on the output of a parser in the form of phrase structure trees or dependencies. There is a great variety of methods for representing sentential meaning (by embedding them in a low dimensional space) which we aim to explore in this project. Moreover, we will investigate novel ways of representing entire documents and non linguistic input. We will also devise various decoders for rendering source input to a probability distribution over target outputs.

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. By deploying these applications on the web, we aim to improve Internet access and enable users to find information that would otherwise be absent or intractable. We will focus on applications that produce textual output whilst translating from linguistic or non-linguistic input. Examples of text-to-text translation include simplification, and summarization-related applications such as the automatic generation of highlights and tweet messages for news articles. With regard to non-linguistic input, we will focus on three problems that translate between language and other modalities: concept-to-text generation, image description generation, and generation of documentation for source code.

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; Woodsend and Lapata, 2010), databases (Barzilay and Lapata, 2006; Konstas and Lapata, 2013; Reddy et al., 2014), and multimodal datasets (Feng and Lapata, 2010; Silberer et al., 2013). The first objective (i.e., formalizing the types of comparable data available and how these can be mined) 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, the development of a unified framework for translating from comparable corpora, is a new and exciting direction that builds on the PI's work on developing robust models and efficient inference mechanisms for natural language. In collaboration with students and colleagues at MIT and Edinburgh, she has previously worked on a variety of translation problems (Barzilay and Lapata, 2008; Clarke and Lapata, 2008; Cohn and Lapata, 2009; Feng and Lapata, 2013; Konstas and Lapata, 2012a) using joint models which can be seen as a pilot for this proposal. Her recent work on neural networks for representing and generating natural language (Silberer and Lapata, 2014; Zhang and Lapata, 2014) can be seen as a pilot for the current proposal. As for the third objective, the PI has developed a variety of applications, including summarization, text simplification, and more recently has begun to work on problems that integrate language and vision (Feng and Lapata, 2013; Ortiz et al., 2015; Silberer and Lapata, 2012).

3. Expected Impact

The impact of this proposed project is wide-ranging, and not limited to the Natural Language Processing Research community. We summarize how our results will affect NLP and a number of related research subareas as well as the general public.

NLP and Related Disciplines 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 ability to translate text in a different style or more concise manner lies at the heart of applications such as summarization, information extraction and question answering. The proposed work is also of direct relevance to information retrieval. For example, it is well known that query reformulation can increase the recall of information retrieval engines (Jones and Tait, 1984). 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).

The algorithmic and modeling developments of this project will be of interest to the Machine Learning and Computer Vision communities and to the related field of SMT. Parallel texts provide indispensable training data for SMT, however they are often limited in size, language coverage, and register. One way to alleviate the lack of parallel data is to exploit comparable corpora which are easier to come by, while more diverse and noisy. Research into translating from non-linguistic input will also strengthen existing work on grounded language acquisition, where the goal is to extract representations of the meaning of natural language tied to the physical world (Branavan et al., 2009; Wong and Mooney, 2007).

The General Public Overall, the proposed research will render the Internet more accessible to a broader audience. The technology developed as part of this project will (a) facilitate the development of reading aids for a wide range of users (e.g., low-literacy readers, non-native speakers), (b) reduce the problem of information overload that has resulted from the proliferation of online data and (c) alleviate the problem of accessing non-textual information on the web for sighted and visually impaired users.

b. Methodology

The proposal is organized into three Work Packages (WPs), each WP dealing with one of the objectives given in Section 2. Although there is a temporal flow from WP1 to WP3, there is also some natural overlap, as it is not possible for example to consider a model or a corpus without having a specific task or application in mind.

4. WP1: The Translation Task

This WP will formalize the translation problem by studying its different manifestations in naturally occurring data. An important prerequisite to the modeling work discussed in WP2 is the collection and analysis of corpus data representative of the breadth, variation, and difficulty of the translation task. To this end, we will classify comparable corpora according to their overlap, compatibility, modality, and so on. A popular way of acquiring a corpus is collecting it from the web, as it provides easy access to an unlimited amount of data. We will focus on monolingual comparable corpora for text-to-text translation tasks and corpora that contain non-linguistic data (e.g., images) and corresponding textual descriptions. In both cases we will examine corpora representative of different domains, genres, and target audiences.

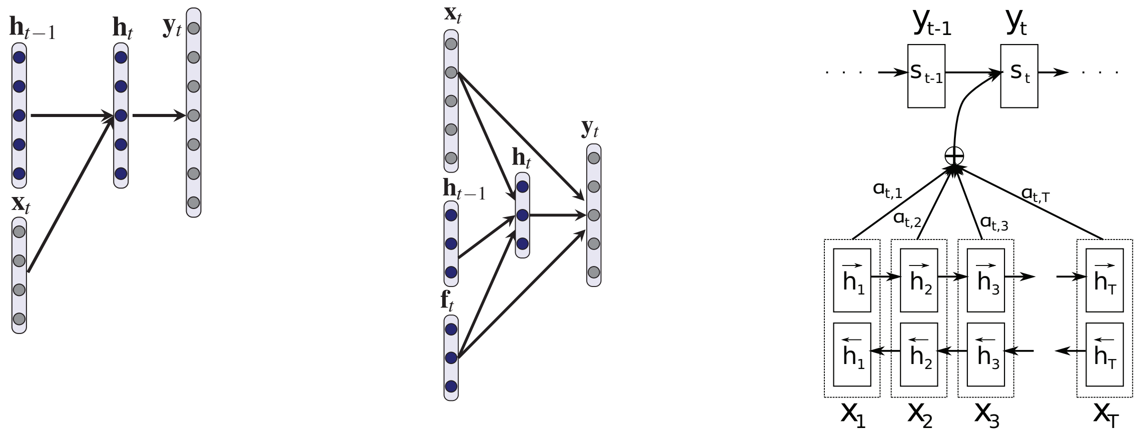
Text-to-Text Translation For text-to-text translation problems, we will consider news articles and their highlights building on the corpus created in Woodsend and Lapata (2010). As mentioned earlier, we will also gather corpora consisting of tweets and the articles they refer to. Wikipedia provides a remarkable resource for the creation of comparable corpora. To name a few possibilities. By mining the Simple English Wikipedia and the main English Wikipedia, we can create a comparable simplification corpus (Woodsend and Lapata, 2011). We could also explore the revision histories, to gather data pertaining to simplification, compression, and general rewriting operations. Another comparable corpus could be constructed from Wikipedia infoboxes and their accompanying articles. Infoboxes are tabular summaries of the most relevant facts contained in an article.

Finally, as a large number of Internet users search the Web for health care information (e.g., to check specific diseases and treatments) we will also construct comparable corpora containing lay and specialized medical texts. For example, a corpus on disease symptoms or treatments can be created by pairing documents from the NHS Direct website with articles from Webster's New World Medical Dictionary. NHS Direct is a health advice and information service that allows patients to check their symptoms if they are feeling unwell. The advice is written in a non-technical style targeting a lay audience.

During WP1 we will identify the types of comparable corpora that exist and how these can be gathered. In some cases, documents can be directly downloaded from existing websites (see the discussion on the Wikipedia corpora above). In other cases, we may need to create the corpus by pairing documents from unrelated websites, whilst making sure that they cover similar topics and are relevant to the genre targeted (e.g., lay versus specialized texts). For example, we may need to classify each document as belonging to one genre or the other. This can be done by automatic categorization of texts or by direct knowledge of the sources of documents. We will also undertake some manual annotation in order to quantitatively measure how comparable the corpora are (e.g., by aligning textual units that express similar content). The annotated data will in addition allow us to evaluate the performance of our automatic methods at identifying segments with shared content (see Section 5 for details).

Translation from Non-linguistic Input Learning how to translate non-linguistic input into text is a relatively new research area. As a result, there are no well-established benchmark datasets and comparisons between systems are few and far between. We aim to rectify this by constructing several database-to-text, image-to-text, and source code-to-text comparable corpora. There are plenty of weather-specific databases on the web. These provide a structured representation of weather conditions (e.g., temperature, sky conditions) and are typically accompanied with short weather reports. In the sports domain, databases and accompanying text can be found for football, cricket, tennis, and many other popular sports. Similar information exists for disastrous events such as earthquakes and aviation accidents. For example, for an earthquake the database will record the year, time, magnitude, latitude/longitude, location, number of deaths, and so on.

Image-to-text corpora can be gathered relatively straightforwardly from news articles (Feng and Lapata, 2013). The latter often contain many images and



(a) Recurrent Neural Network

(b) Recurrent neural network model, including auxiliary input layer $\mathbf{f}_t$.

(c) Recurrent neural network with bidirectional RNN encoder

Figure 1: Examples of different neural network model architectures.

their associated captions. Aside from news, images with naturally occurring descriptions can be obtained from Wikipedia, where an image may be used in several articles with semantically related captions (see Table 4) and the Flickr photo sharing site (Ordonez et al., 2011). Other interesting datasets include images found on shopping websites together with product descriptions, famous landmarks and short summaries of their features and historic importance, or medical images accompanied with explanations about the disease and its symptoms.

Billions of lines of code are readily available on the Internet, much of which are of professional quality and well-documented. Furthermore, there exist source code documentation templates for several programming languages that provide a standard for where and how to create documentation. For example, Java programs are usually documented using the javadoc tool which collates all the comments from a series of Java files and produces an overview of all the packages, classes and their methods. This will allow us to collect comparable corpora of source code and its documentation relatively easily. Besides Java, Python is another programming language with standardized documentation standards. In addition, programming language courses are an excellent source for gathering examples of documented source code as well as Question and Answer (Q&A) sites such as StackOverflow[2] which contain a wealth of a NL descriptions and their corresponding code segments.

Analogously to the text-to-text case, we will also create corpora covering different domains (for databases and images) and programming languages and annotate the semantic correspondences between the source and the target. For the image-to-text datasets we will also label the objects and events depicted and how they align to words, phrases or larger syntactic units. Although we ultimately aim at inducing this latent information automatically, we will obtain gold-standard data to gauge the difficulty of the task, to allow for controlled experiments (e.g., where one source of noisy information is partialled out) and for evaluation purposes. We anticipate to use a modified version of LabelMe (Russel et al., 2005), a web-based annotation tool that allows easy image annotation and instant sharing of such annotations. The online tool provides functionalities such as drawing polygons, querying images, and browsing the annotation database. We will collect annotations throughout the course of the project, thereby creating a large scale dataset for the development and testing of our translation models.

5. WP2: Modeling Framework

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

Network Architecture Figure 1a illustrates a typical recurrent neural network (RNN) architecture containing three layers, an input layer, a hidden layer, and an output layer. The input layer is a concatenation of $\mathbf{h}_{t-1}$ and $\mathbf{x}_t$, where $\mathbf{h}_{t-1}$ is a real-valued vector, encoding the history of all words observed in the sequence up to time step $\mathbf{t}-1$. $\mathbf{x}_t$ is the embedding of the input word at time t. 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

[2] http://stackoverflow.com/

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 major attraction of recurrent architectures is their potential to capture long-span dependencies since predictions are based on an *unbounded history* of previous words. This is in contrast to feed-forward networks as well as conventional n-gram models, both of which are limited to fixed-length contexts. Nevertheless, 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 modeling the conditional dependence of the output text given some input. A straightforward way to model the translation task is to add a continuous space representation of the source as an additional input 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 $\mathbf{f}_t$ in Figure 1b can be used to feed in arbitrary additional information, however here we focus on encodings 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 order to perform translation, 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 the network architecture 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, and a subset of these is chosen adaptively while decoding the input. 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* $h_1 \dots h_T$ to which the encoder maps the source input. The annotations summarize the information found in preceding *and* following words. Furthermore annotations have weights expressing which parts of the input are important in generating the target output.

The architectures in Figures 1 specify a general modeling framework; depending on how the encoder is defined and how the decoder estimates the conditional probability $p(\mathbf{x}|\mathbf{y})$ a variety of models can be expressed. We give examples of model candidates we plan to develop in this project below.

Encoder Layer $\mathbf{h}_t$ in Figure 1 encapsulates information about the meaning of the source sentence, whereas $\mathbf{f}_t$ represents auxiliary information (and is also sentence-based). Various neural sentence models have been described in the literature. A general class of basic sentence models consists of a projection 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 (Mikolov et al., 2010; Socher et al., 2011; Turian et al., 2010).

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 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. As shown in Figure 2a, $s^{[l,m]}$ and $s^{[m,n]}$ are the representations of child nodes; $s^{[l,n]}$ 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). They could be also encode documents, e.g., by using the output of a text-level discourse parser (Feng and Hirst, 2014).

A third class of encoders is based on the convolution operation (Collobert et al., 2011; 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 sentences. 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 for a seven word sentence in shown in Figure 2b. The model takes as input a sentence matrix whose columns correspond to vectors each representing a word in a sentence and performs a series of one dimensional convolutions. The latter are applied feature-wise, i.e., across each feature of the word vectors in the sentence. 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,

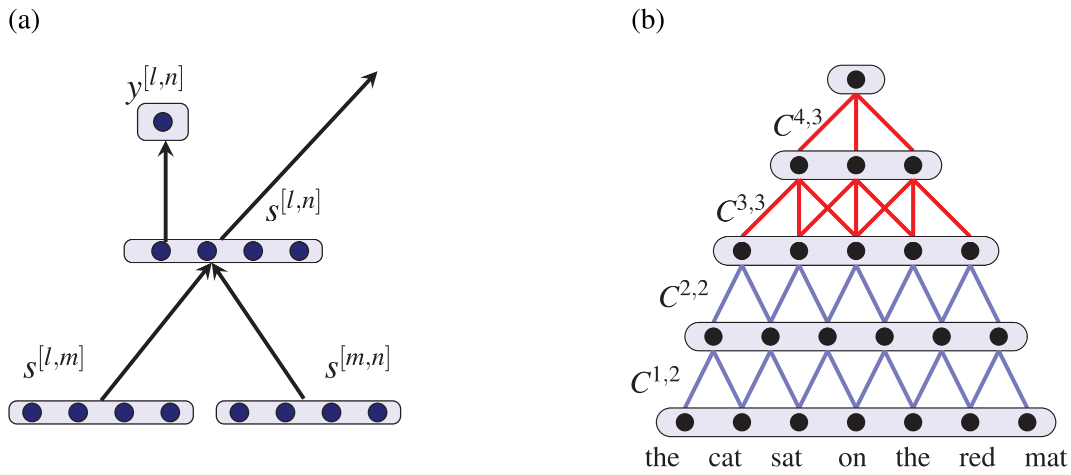


Figure 2: Examples of encoders: (a) recursive neural network; (b) convolutional sentence model (the first layer has seven vectors, one for each word. Two neighboring vectors are merged to one vector in the second layer with weight matrix $C^{1,2}$. In other layers, either two or three neighboring vectors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

deep Convolutional neural networks are nowadays the architecture of choice for large-scale image recognition (Krizhevsky et al., 2012; 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 (Auli et al., 2013; Cho et al., 2014; 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., verb phrases, noun phrases, prepositional 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 (e.g., Woodsend and Lapata 2011). 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.

6. WP3: Development of Applications

We will use several applications as a means of evaluating our framework and showcasing its practical utility. In the following, we first discuss text-to-text applications and then move on to applications that translate from non-linguistic input.

Text-to-Text Translation Text simplification is perhaps one of the oldest text-to-text translation problems. Given a source document, the goal is to create a grammatical target that is easier to read with simpler vocabulary and syntactic structure (see example in Table 1).

Individuals with low literacy stand to benefit the most from such an application. It is well known that literacy skills impact every aspect of adult life including the use of the Internet (Summers and Summers, 2005). For instance, lower-literacy users cannot read a text by glancing it. They must read word for word and often spend considerable time trying to understand multi-syllabic words. When confronted with long, dense web pages, complex syntax, unfamiliar words, and parenthetical text, lower-literacy users simply skip entire paragraphs. They also have difficulty processing search results, which typically consist of out-of-context text snippets. As a result, they often simply pick the first hit on the list, even if it's not the most appropriate for their needs. From an engineering perspective, a simplification component could be also used as a preprocessing step to improve the performance of parsers, summarizers or machine translation engines.

Aside from text simplification, we will also focus on summarization for two reasons. Firstly, summarization could help users digest the large amounts of information available on the web. Secondly, the task is a prime example of the challenges faced by the translation problems considered in this proposal. Summaries must be maximally informative, minimally redundant, grammatical, coherent, and adhere to a prespecified length. We will apply our models to the task of generating highlights for a single document. Highlights give a brief overview of the article to allow readers to quickly gather information on stories, and usually appear as bullet points. Importantly, they represent the gist of the *entire* document and thus often differ substantially from the first couple of sentences in the article. They are also highly compressed, written in a telegraphic style and thus provide an excellent testbed for models that perform rewriting operations. Aside from casual Internet users reading news, the proposed application stands to benefit commercial news agencies and networks alike who typically do not include story highlights in their articles. A notable exception is CNN, who nevertheless creates highlights manually.

We will also address the task of tweet generation for news articles. Tweets are text messages up to 140 characters. More than 50 million tweets are generated daily by millions of Twitter users, a substantial portion of which are related to news according to a study conducted by Pear Analytics (2009). In fact, many online news sites (e.g., BBC, CNN, the Wall Street Journal) have their own twitter feeds where journalists post short messages relating to their articles together with links to the full text (see Table 2). These tweets are typically well-written, include few abbreviations and are good examples of extreme summarization. Interestingly, journalists do not simply cut and copy sentences from the main article; rather they summarize its gist whilst trying to pack as much information as possible given the 140 characters constraint. Although we propose to generate tweets automatically for news articles, we believe that such an application is not necessarily news specific and could be deployed for the weather, scientific articles, movie reviews, blogs, etc.

Translation from Non-linguistic Input While existing concept-to-text generation systems can be engineered to obtain good performance on particular domains (e.g., Turner et al. 2009), it is often difficult to adapt them across different domains as they rely mostly on handcrafted components. Our goal is to build a simple and flexible system which is domain-independent and portable. As explained earlier, we will reduce content selection and surface realization into representation learning problem, i.e., by appropriately encoding the structure of the database and its correspondence to natural language. Much information on the web is in a structured but non-linguistic form, and as a result currently inaccessible to search engines. Using a system like the one advocated here, would allow to search for and retrieve non-linguistic data via its translation into text. We will develop translation applications for three domains, i.e., the weather, sports, and accidents (e.g., earthquakes).

As mentioned earlier, we are not aware of any existing systems that automatically generate documentation for source code. Such a tool could be used by developers to semi-automatically document their code, by search engines in order to retrieve code snippets that exemplify a method or function, and as teaching aids for learners of programming languages. It is important to note that our tool will be different from current documentation generators. The latter are programming tools that generate software documentation intended for programmers or end users from a set of specially commented source code files. The comments have been manually written and are already available, whereas we intend to supply the comments automatically from scratch. We will develop a code-to-text system for Java, an object orientated programming language which is popular amongst programmers and for which large online code depositories are available online.

Finally, we will also develop an image description generation application. By description we mean text consisting of one or more sentences rather than a few isolated keywords. One reason for using more linguistically meaningful descriptions is that keywords are often ambiguous. An image annotated with the words *blue*, *sky*, *car* could depict a blue car or a blue sky, whereas the caption *car racing under the blue sky* would make the relations between the words explicit. Automatic description generation could therefore improve image retrieval by supporting longer and more targeted queries. It could also assist journalists in creating captions for the images associated with their articles and increase the accessibility of the web for visually impaired users who cannot access the content of many sites in the same ways as sighted users can. The recent success of Convolutional Neural Networks on a variety of computer vision tasks in the last few years has sparked a great deal of enthusiasm about the image description generation task. Many approaches use the representation of images generated by CNN, trained for object recognition tasks as a starting point (Krizhevsky et al., 2012; Simonyan and Zisserman, 2014). We will experiment with CNN-based representations as well as with encodings based on bounding boxes and visual dependencies (Ortiz et al., 2015).

The description generation task poses interesting questions. Can we learn a representation that captures image-to-word correspondences? Can we analyze images in a hierarchical fashion so as to exploit the structural parallelism between the textual and the visual modalities? We will generate descriptions for general images (e.g., those taken from the news) and more specialized ones (e.g., those found in different Wikipedia categories or consumer sites) as well as publicly available datasets (e.g., Flickr 30K, COCO).

Evaluation All the tasks discussed in this proposal will be subject to the same evaluation protocol. We will employ automatic evaluation measures during system development, whereas stable system versions will be assessed with real users. Luckily, a variety of automatic evaluation metrics have been proposed for rewriting tasks. Summarization applications typically use ROUGE, a task specific metric that measures the similarity between system and reference summaries (Lin and Hovy, 2003). Other applications such as simplification and image caption generation, have used the well-known BLEU score (Papineni et al., 2002). In addition, the output of a simplification system can be assessed using one of the many existing readability measures such as the Fleisch-Kincaid Grade Level index (see Mitchell 1985 for an overview).

As automatic evaluation metrics are mostly based on shallow text features (e.g., word overlap, word and syllable length), they not ideal for assessing how users perceive the translations. In addition to automatic evaluation, we will conduct several human studies to assess system performance on different dimensions, e.g., correctness, grammaticality, coherence. As an example consider the simplification case. A hypothetical model can be evaluated with high and low-literacy readers. Participants will read documents and then answer comprehension questions. They will be given various textual treatments: the original text, a system-produced simplified version, a human-authored simplified version. Participants will rate the output in terms of whether it is grammatical, simpler than the source document and truth preserving. In all cases, system evaluation will use benchmark test sets if these are available. We will conduct similar studies for our other applications.

In all cases we will build on-line versions of the software with functionality similar to Google translate where the user can key in a document (or URL), image, source code to obtain a simpler version, a summary, a caption, or documentation. Together with the output, users will be provided a short questionnaire on assessing its quality. This way, we will obtain direct feedback on our software and large amounts of evaluation data for improving system output.

7. Expected Outputs

The expected outputs of the proposed project will take on a variety of forms:

Peer-reviewed publications The principal dissemination mechanism will be publications in peer-reviewed journals and conferences. Aside from well-known NLP venues, the proposed research would benefit from exposure to other research communities such as Machine Learning and Computer vision. Suitable journals include Computational Linguistics, the Journal of Artificial Intelligence Research, Journal of Machine Learning Research, the IEEE Transactions on Pattern Analysis and Machine Intelligence. Besides ACL and associated conferences (e.g., NAACL, EACL) other related venues include NIPS, ICML, ICCV, CVPR, and AAAI.

Working Implementations Software and documentation will be made publicly available for research purposes. We will also make our training and testing corpora freely available to the community in order to create standard evaluation benchmarks. A website will be developed for distributing information, data and code, and hosting web demonstrations. The PI already has a track record of producing software and resources used by the international research community, and releasing the software from this project will ensure maximum exposure and uptake.

Workshops The PI will organize (at least) two workshops, one mid-way through the project at one of the ACL conferences, aimed at the language processing community. Related workshops have been held previously,[3] demonstrating the interest from the field in the general research problem. The second workshop will be more inter-disciplinary and will be held at a non-NLP venue (e.g., NIPS, ICCV, CVPR).

8. Research Environment

The project will be located in the School of Informatics at the University of Edinburgh. The School was the highest-ranking department in the area of computer science and informatics in the UK in the last two Research Assessment Exercises (2008 and 2014). In the most recent exercise, Informatics was assessed as delivering more world-leading and internationally excellent (4* and 4*) research than any other university in the UK. The School is located in the new Informatics Forum, a purpose-built hub for informatics in Edinburgh's city center housing over 500 researchers.

Within Informatics, the proposed project will be based in ILCC, the Institute for Language, Cognition and Computation. ILCC is one of seven institutes in Informatics, comprising about 100 researchers. It conducts research on all aspects of natural language processing, drawing on machine learning, statistical modeling, and computational, psychological, and linguistic theories of communication among humans and between humans and machines, using text, speech and other modalities. Together with ILCC colleagues Sharon Goldwater and Frank Keller, the PI is running the Probabilistic Models of Language research group. This is a weekly meeting of staff and students interested in probabilistic and statistical methods, as used in both natural language processing systems and models of human language acquisition and processing. It involves informal paper discussions and more formal research presentations, as well as conference practice talks and other research-related activities.

[3] A Workshop on Building and Using Comparable Corpora was held at ACL-HLT 2011; A workshop on Semantic Interpretation in an Actionable Context took place at NAACL 2012; and workshop on integrating language and vision were organized at NIPS 2011, NAACL 2013, and ECCV 2014.

Cost Category			**Total in Euro**
Direct Costs	**Personnel**	PI	€ 477,532
		Senior Staff	€ -
		Postdocs	€ 447,911
		Students	€ 308,392
		Other	€ 80,453
	i. Total Direct costs for Personnel (in Euro)		*€ 1,314,288*
	Travel		€ 91,785
	Equipment		€ 72,500
	Other goods and services	Consumables	€ 36,250
		Publications (including Open Access fees), etc.	€ -
		Other (please specify) Audit	€ 5,800
	ii. Total Other Direct Costs (in Euro)		*€ 206,335*
A – Total Direct Costs (i + ii) (in Euro)			**€ 1,520,623**
B – Indirect Costs (overheads) 25% of Direct Costs] (in Euro)			**€ 380,156**
C1 – Subcontracting Costs (no overheads) (in Euro)			€ -
C2 – Other Direct Costs with no overheads (in Euro)			€ -
Total Estimated Eligible Costs (A + B + C) (in Euro)			**€ 1,900,779**
Total Requested EU Contribution (in Euro)			**€ 1,900,779**

For the above cost table, please indicate the % of working time the PI dedicates to the project over the period of the grant:	**70%**

Table 5: Project costs (in euros)

The PI also maintains several collaborations with colleagues at ANC, the Institute for Adaptive and Neural Computation and IPAB, the Institute of Perception Action and Behavior. She has been co-supervising students with Dr Charles Sutton (ANC) and Dr Vittorio Ferrari (IPAB) since 2009. Their expertise in machine learning and computer vision, respectively, will be beneficial for this project (see the Resources section for details). Dr Sutton works on inference methods for graphical models and is interested in the application of statistical techniques from natural language to programming language text. Dr Ferrari works on unsupervised methods for computer vision (e.g., object recognition, pose estimation, image segmentation) and is collaborating with the PI on developing novel visually grounded meaning representation models. The PI will also work with Prof Philip Wadler (LFCS) on the development of code-to-text translation models. Prof Wadler's research interests are in programming languages; he is one of the main contributors to the design of Haskell.

c. Resources

Size and Nature of the Team The proposed project requires a team that includes both language processing and machine learning experts. The team also needs to be methodologically diverse, with expertise in neural networks, algorithms and data structures, image processing. The principal investigator 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 (i.e., creation of comparable corpora, formulation and implementation of the translation framework, applications). In addition, she will provide scientific leadership and management throughout the project.

The research team will also include two postdoctoral researchers, with expertise in NLP, computer science and machine learning. 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 primarily responsible for developing translation models from non-linguistic data. It is envisaged that 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 (co-supervised with Dr Ferrari) will focus on the image description generation application and the other one (co-supervised with Prof Sutton) 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. We expect project team members to attend two to three conferences per year in order to ensure appropriate dissemination across different fields. For example, the conferences could include one computer vision conference, one NLP conference, and one machine learning conference per year. The PI will also make three extended visits to the US. These involve spending three months at Microsoft Research Redmond (Bill Dolan's group) and three months at Stanford (Fei Fei Li's lab) and three months at Google Mountain View. The visits will be instrumental in publicizing the work undertaken in this project and also to initiate fruitful collaborations. Table 5 gives a detailed breakdown of the project costs.

Infrastructure and equipment The proposed work is highly data-intensive, and thus five GPU servers are requested to build a large-scale computing infrastructure for experimentation. Funds for a laptop are also requested to support dissemination of the project findings at conferences and for providing demos of the developed systems and software. As the project involves processing of large amounts of textual, image, and other data and the building of large-scale inference models, these unusual computer resources are deemed necessary.

References

Angeli, G., Liang, P., and Klein, D. (2010). A simple domain-independent probabilistic approach to generation. In *Proceedings of the 2010 EMNLP*, pages 502–512, Cambridge, MA.

Auli, M., Galley, M., Quirk, C., and Zweig, G. (2013). Joint language and translation modeling with recurrent neural networks. In *Proceedings of the 2013 EMNLP*, pages 1044–1054, Seattle, Washington, USA.

Bach, N., Gao, Q., Vogel, S., and Waibel, A. (2011). TriS: A statistical sentence simplifier with log-linear models and margin-based discriminative training. In *Proceedings of 5th IJCNLP*, pages 474–482, Chiang Mai, Thailand.

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

Barzilay, R. (2003). *Information Fusion for Multi-Document Summarization: Paraphrasing and Generation.* PhD thesis, Columbia University.

Barzilay, R. and Lapata, M. (2005). Collective content selection for concept-to-text generation. In *Proceedings of HLT-EMNLP*, pages 331–338, Vancouver, British Columbia, Canada.

Barzilay, R. and Lapata, M. (2006). Aggregation via set partitioning for natural language generation. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 359–366, New York City, USA. Association for Computational Linguistics.

Barzilay, R. and Lapata, M. (2008). Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34:1–34.

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

Bengio, Y., Ducharme, R., Vincent, P., and Janvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, (1137–1155).

Branavan, S., Chen, H., Zettlemoyer, L., and Barzilay, R. (2009). Reinforcement learning for mapping instructions to actions. In *Proceedings of the 47th ACL the 4th IJC-NLP*, pages 82–90, Suntec, Singapore.

Carroll, J., Minnen, G., Pearce, D., Canning, Y., Devlin, S., and Tait, J. (1999). Simplifying text for language impaired readers. In *Proceedings of the 9th EACL*, pages 269–270, Bergen, Norway.

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.

Clarke, J. and Lapata, M. (2008). Global inference for sentence compression: An integer linear programming approach. *Journal of Artificial Intelligence Research*, 31:273–381.

Cohn, T. and Lapata, M. (2009). Sentence compression as tree transduction. *Journal of Artificial Intelligence Research*, 34:637–674.

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.

Duboue, P. A. and McKeown, K. R. (2002). Content planner construction via evolutionary algorithms and a corpus-based fitness function. In *Proceedings of International Natural Language Generation*, pages 89–96, Ramapo Mountains, NY.

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. (2010). How many words is a picture worth? automatic caption generation for news images. In *Proceedings of the 48th ACL*, pages 1239–1249, Uppsala, Sweden.

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

Ganitkevitch, J., Callison-Burch, C., Napoles, C., and Van Durme, B. (2011). Learning sentential paraphrases from bilingual parallel corpora for text-to-text generation. In *Proceedings of the 2011 EMNLP*, pages 1168–1179, Edinburgh, Scotland, UK.

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

Hodosh, M., Young, P., and Hockenmaier, J. (2013). Framing image description as a ranking task: Data, models and evaluation metrics. *JAIR*, 47.

Jing, H. (2000). Sentence reduction for automatic text summarization. In Nirenburg, S., editor, *Proceedings of the 6th ANLP*, pages 310–315.

Jones, K. S. and Tait, J. I. (1984). Automatic search term variant generation. *Journal of Documentation*, 40(1):50–66.

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

Knight, K. and Marcu, D. (2002). Summarization beyond sentence extraction: a probabilistic approach to sentence compression. *Artificial Intelligence*, 139(1):91–107.

Koehn, P. (2009). *Statistical Machine Translation*. Cambridge University Press, Cambridge.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of the 2003 HLT*, pages 48–54, Edmonton, Canada.

Konstas, I. and Lapata, M. (2012a). Concept-to-text generation via discriminative reranking. In *Proceedings of the 50th ACL*, pages 369–378, Jeju Island, Korea.

Konstas, I. and Lapata, M. (2012b). Unsupervised concept-to-text generation with hypergraphs. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 752–761, Montréal, Canada.

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

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1106–1114.

Kulkarni, G., Premraj, V., Dhar, S., Li, S., Choi, Y., Berg, A. C., and Berg, T. L. (2011). Baby talk: Understanding and generating image descriptions. In *Proceedings of the 24th CVPR*, Colorado Springs, CO.

Lin, C.-Y. and Hovy, E. H. (2003). Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of HLT–NAACL*, pages 71–78, Edmonton, Canada.

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.

Mitchell, J. V. (1985). *The Ninth Mental Measurements Year-book*. University of Nebraska Press, Lincoln, Nebraska.

Mitchell, M., Dodge, J., Goyal, A., Yamaguchi, K., Stratos, K., Mensch, A., Berg, A., Berg, T., and Daum, H. (2012). Midge : Generating Image Descriptions From Computer Vision Detections. In *EACL*.

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.

Och, F. J. (2003). Minimum error rate training in statistical machine translation. In *Proceedings of the 41st ACL*, pages 160–167, Sapporo, Japan.

Ordonez, V., Kulkarni, G., and Berg, T. L. (2011). Im2text: Describing images using 1 million captioned photographs. In *Advances in Neural Information Processing Systems*, pages 1143–1151, Granada, Spain.

Ortiz, L. G. M., Wolff, C., and Lapata, M. (2015). Learning to interpret and describe abstract scenes. In *Proceedings of NAACL*, Denver, Colorado. to appear.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th ACL*, pages 311–318, Philadelphia, PA.

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.

Russel, B. C., Torralba, A., Murphy, K. P., and Freeman, W. T. (2005). Labelme: A database and web-based tool for image annotation. Aim-2005-025, MIT AI Lab Memo.

Siddharthan, A. (2006). Syntactic simplification and text cohesion. *Research on Language and Computation*, 4(1):77–109.

Silberer, C., Ferrari, V., and Lapata, M. (2013). Models of semantic representation with visual attributes. In *Proceedings of the 51st ACL*, pages 572–582, Sofia, Bulgaria.

Silberer, C. and Lapata, M. (2012). Grounded models of semantic representation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1423–1433, Jeju Island, Korea.

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.

Socher, R., Lin, C. C., Ng, A. Y., and Manning, C. D. (2011). Parsing natural scenes and natural language with recursive neural networks. In *Proceedings of the 26th International Conference on Machine Learning*, Bellevue, Washington, USA.

Summers, K. and Summers, M. (2005). Reading and navigational strategies of web users with lower literacy skills. In *Proceedings of the 68th Annual Meeting of the American Society for Information Science ad Technology*, Charlotte.

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

Turian, J., Ratinov, L.-A., and Bengio, Y. (2010). Word representations: A simple and general method for semisupervised learning. In *Proceedings of the 48th ACL*, pages 384–394, Uppsala, Sweden.

Turner, R., Sripada, Y., and Reiter, E. (2009). Generating approximate geographic descriptions. In *Proceedings of the 12th European Workshop on Natural Language Generation*, pages 42–49, Athens, Greece.

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

Wong, Y. W. and Mooney, R. (2007). Learning synchronous grammars for semantic parsing with lambda calculus. In *Proceedings of the 45th ACL*, pages 960–967, Prague, Czech Republic.

Woodsend, K. and Lapata, M. (2010). Automatic generation of story highlights. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 565–574, Uppsala, Sweden. Association for Computational Linguistics.

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.

Woodsend, K. and Lapata, M. (2012). Multiple aspect summarization using integer linear programming. In *Proceedings of the 2012 EMNLP-CoNLL*, pages 233–243, Jeju Island, Korea.

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.

Zhu, Z., Bernhard, D., and Gurevych, I. (2010). A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd COLING*, pages 1353–1361, Beijing, China.