IDIAP RESEARCH REPORT

RESEARCH INSTITUTE

TOWARDS DOCUMENT-LEVEL NEURAL MACHINE TRANSLATION

Lesly Miculicich Werlen

Idiap-RR-25-2017

SEPTEMBER 2017

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

Research proposal submitted to
the Doctoral Program in Electrical Engineering

Towards Document-Level Neural Machine Translation

PhD Student	Lesly S. Miculicich Werlen
President of the Commitee	Prof. Sabine Süsstrunk
Expert	Dr. Martin Rajman
Thesis Director	Prof. Hervé Bourlard
Co-director	Dr. Andrei Popescu-Belis

September 14, 2017

Contents

Summary

Machine Translation (MT) has made considerable progress in the past two decades, particularly in the last few years due to the introduction of architectures based on neural networks (NNs). Before NNs, the state-of-the-art was dominated by statistical approaches, which, despite the extensive research work, were surpassed by the power of NNs. MT models learn to translate from large amounts of parallel sentences in different languages. The focus on sentences brings a practical simplification for the task, but also has the disadvantage of missing contextual information while translating a document. This issue cannot be minimized, as MT is commonly used for translating entire documents rather than individual sentences. One key point is that the dependencies among distant words, related to discourse-level constraints, are being ignored, even inside a single sentence. This often results in a lack of coherence in the translated text. Although several research studies have approached this problem in the framework of statistical machine translation, to the best of our knowledge, no research work has been done yet in the framework of neural machine translation (NMT) on discourse-level constraints.

The objective of this thesis is to improve the automatic translation of documents by including discourse-level constraints. We address the problem in four stages. First, we will focus on the connections at the sentence level. Currently, sentences are modeled as sequences, where each word is conditioned on the previous ones. However, in practice, this model does not fully capture the long-range dependencies between words, in particular, those related to lexical choice of entity mentions (i.e. nouns or pronouns). We propose to enhance the sequential model to infer long-term connections. For this purpose, we plan to extend its memory capacity, and incorporate an “attention” mechanism to help the network to focus in specific part of the sentence useful to predict a particular word. Second, we will expand this idea to model connections among different sentences. Inspired by the notion of human reading, we propose to incorporate an external “memory” to give the network the capacity to retain information from past sentences, and to create internal representations of a text. Third, we will formulate and study the problem of document-level neural MT. The idea is to model and optimize the translation of a document as a whole (instead of independent sentences), including the same internal connections among words as above. Finally, we consider the possibility to incorporate, in the document-level architecture, internal mechanisms such as coreference resolution to enforce the learning of discourse-level dependencies by a multitask learning approach.

In summary, the contributions of the thesis will be as follows:

(i) Incorporating long-term dependencies for modeling discourse constraints at sentence-level NMT.
(ii) Modeling memory representations of a text to enhance the capacity of the network to make simple reasoning for text understanding.
(iii) Proposing a new neural network architecture for document-level NMT.
(iv) Learning discourse constraints from annotated data for document-level NMT.

During the past year, we addressed the specific problem of improving the translation of entity mentions with the help of coreference resolution (i.e. grouping mentions that refer to the same entity). We implemented a coreference-aware translation system that helps to disambiguate the translation of mentions by optimizing the similarity of mention-grouping in source and target documents. We designed two different approaches: the first one selects the set of sentences which maximize a document-level similarity score; the second one selects, for each entity, the set of mentions which maximize a cluster-level similarity score. We tested both approaches on Spanish-English translation, and obtained significant improvements in the translation of pronominal mentions compared to the baseline.

The research proposal is organized as follows: Section 1 reviews the state-of-the-art in MT, Section 2 presents the research performed during the first year, Section 3 offers a detailed description of the research plan for the following years; and finally Section 4 contains the timetable.

1 State-of-the-art

1.1 Statistical Machine Translation

Statistical machine translation (SMT) (Brown et al., 1990) is a machine translation (MT) approach based on statistical models. The parameters of the model are learned from a corpus of parallel sentences in source and target languages. A sentence is represented as a sequence of words, or tokens (including punctuation, and special treatment for compound words), and the objective is to find the most probable sentence in the target language e_{best} given a sentence in the source language f:

$$e_{best} = \arg\max_{e} p(e|f) \approx \arg\max_{e} p(f|e)p(e)$$

The translation is expressed as an optimization problem, but the search space increases exponentially according the length of the sentence. Therefore, by applying Bayes Theorem, the problem is divided in two parts: the translation model $p(f|e)$, and the language model $p(e)$. This transformation permits the reduction of the search space by giving priority to well-formed sentences, i.e. with higher probabilities according to the language model. The introduction of the phrase-based translation model (Koehn et al., 2003), and later of the enhanced hierarchical phrase-based translation model (Chiang, 2005) was a big step towards making SMT feasible. The input sentence is divided into "phrases" which are translated independently and then reordered. This approach was the state-of-the-art until very recently, surpassed only by the introduction of neural networks.

The most common evaluation method for MT is the BLEU score (Papineni et al., 2002). It is a simple automatic metric that varies from 0 to 100, and measures the degree of similarity of the translation with one or more reference human translations by calculating the *n-gram* precision between them. Another similar metric is METEOR (Lavie and Denkowski, 2009), which measures *unigram* precision and recall, and utilizes flexible word matching with morphological variants and synonyms.

1.2 Neural Machine Translation

The earliest attempt to use neural networks in MT was to replace the frequency-based Bayesian or *n-gram* language model (Jelinek, 1980) for a neural language model (Bengio et al., 2003; Schwenk et al., 2006). Later, feed-forward neural networks were used to enhance the phrase-based systems by rescoring the translation probability of phrases (Devlin et al., 2014). In spite of improving the translation performance, they required the phrases to have a fixed length. This issue was addressed by recurrent neural networks. The state-of-the-art for MT is based on this arquictecture.

1.2.1 Recurrent Neural Networks

Recurrent neural networks (RNNs) offer a principled way to represent sequences. It estimates a conditional probability distribution of a sequence $x = (x_1, ..., x_T)$ by learning to predict one element at a time given the previous ones $p(x_t|x_1, ..., x_{t-1})$. This is possible because its hidden states form a directed cycle, which allows the network to have dynamic temporal behavior (Figure 1). At time t, the hidden layer is updated as follows: $h_t = f(h_{t-1}, x_t)$, where f is a non-linear activation function. And, the output is emitted according: $y_t = g(h_t)$, where g is a normalization function, usually $softmax$ over the vocabulary.

RNNs, however, suffer of vanishing gradient problem (Pascanu et al., 2013). During the learning process, neurons in early layers learn much more slowly than neurons in later layers. This means that RNNs have a limited memory over time steps. Long short term memory (LSTM) networks (Hochreiter and Schmidhuber, 1997) helps to overcome this problem. LSTMs are RNNs with the capability of maintaining long-term dependencies, thanks to the introduction of specialized

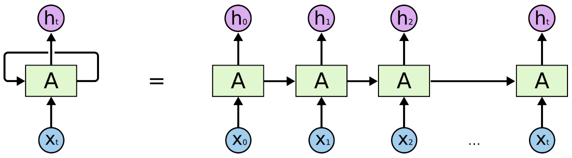


Figure 1: Recurrent neural network: Unfolded representation (Olah, 2015)

gates which allow it to remember and forget information according the learning needs. Current neural machine translation (NMT) systems are based on LSTM or similar arquitectures like gated recurrent units (GRU) (Cho et al., 2014; Chung et al., 2014).

1.2.2 Sequence to Sequence Model

The sequence to sequence model represents a conditional distribution of emitting a sequence given another $p(y_1, ..., y_n | x_1, ..., x_m)$. The initial model applied for machine translation (Cho et al., 2014), consists of two RNNs: an encoder and a decoder. The encoder is a simple RNN that emits an output c only in the last step of the encoding. This output c can be understood as the vector representation of the source sentence in a continuous space, where the assumption is that "similar" sentences are close to each other. Based on this representation, the decoder generates a sequence of words in the target language as follows:

$$p(y_t | y_1, ...y_{t-1}, c) = g(s_t, y_{t-1}, c) \quad s.t. \quad s_t = f(s_{t-1}, y_{t-1}, c) \tag{1}$$

where s_t is the hidden state of the decoder at time t. The conditioning over previously emitted words ensures that the output sequence is a valid sentence, in the same manner as a language model. One drawback of this model is that while short sentences can be represented in a vector successfully, longer ones may not (Pouget-Abadie et al., 2014). The performance of translation degrades with the length of the sentences.

1.2.3 Attention Mechanism

To avoid the degradation of the translation of long sentences, (Bahdanau et al., 2014) introduced an attention mechanism, which allows the decoder to select at each step which part of the source sentence is more useful to predict the next output symbol. Therefore, instead of a unique sentence representation, the output will be dependent on a context vector c_t, which is a weighted sum over all hidden states of the encoder. Equation 1 is modified as follows:

$$\begin{aligned} p(y_t | y_1, ...y_{t-1}, c_t) = g(s_t, y_{t-1}, c_t); \quad s.t \quad s_t = f(s_{t-1}, y_{t-1}, c_t), \quad c_t = \sum_{j=1}^{Tx} \alpha_j^t h_j \\ and \quad \alpha_j^t = softmax(a(s_{t-1}, h_j)) \end{aligned} \tag{2}$$

where h_j is the hidden state of the encoder position j, and α_j^t is a weight calculated from a normalized "alignment function" a, that scores how good is the match between the input at position j and the output at position t. Figure 2 shows an example of translation with attention mechanism where the values of the alignment weights α are represented by the thickness of the lines connecting the source and target sentences. The translation is made from English to German and French. We can see that the alignments change for different languages following their respective grammatical order.

Several other modifications to the attention model have been studied in search of more efficiency, but the performance of translation has not been significantly improved (Luong et al., 2015a; Xu et al., 2015). One advantage of attention-based NMT is that it incorporates flexibility while

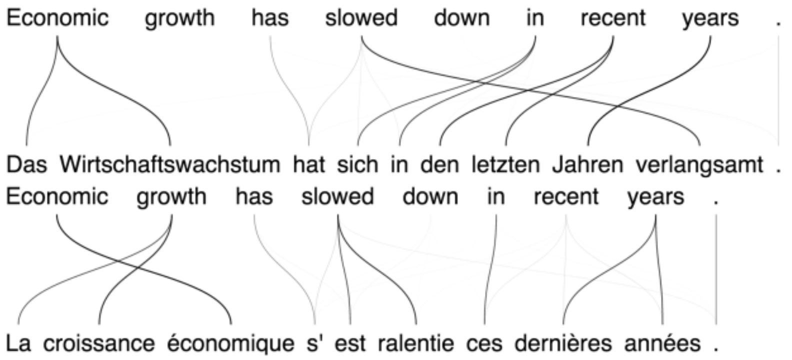

Economic growth has slowed down in recent years .
Das Wirtschaftswachstum hat sich in den letzten Jahren verlangsamt .
Economic growth has slowed down in recent years .
La croissance économique s' est ralentie ces dernières années .

Figure 2: Example of translation using attention mechanism for English to German, and French. (Cho, 2015).

translating language pairs with different grammatical order, a point with which SMT struggles. A comparison between phrase-based SMT and NMT in terms of quality of translation (Bentivogli et al., 2016), concludes that NMT generates fewer morphological, lexical, and word order errors. For instance in the latter case, it produces 70% less errors than SMT. However, the NMT approach suffers for a limited size vocabulary, and it fails to translate all source words, so for example negative clauses are sometimes translated as positive ones (Bentivogli et al., 2016) a type of error which is difficult to recover by a reader with no understanding of the source text.

Some additional techniques to improve the quality of translation of NMT systems are important to finally out-perform SMT (Bojar et al., 2016). Usually in NMT, the words are inputted as one-hot vector. This vector has a limited size (vocabulary size) therefore only the more frequent words are represented. One solution consist on utilize sub-word units to include the modeling of infrequent or rare words (Sennrich et al., 2015). Sub-word units are learned in advance from training data by an algorithm based on byte pair encoding (BPE). Basically, it extracts the more frequent sequences of characters including complete words. A word is represented with the minimum number sub-words that form it. Additinally, bidirectional RNNs are being used to incorporate additional context into word representations. It formulates the dependency of a word not only on the previous words but also on the next ones. The simplest technique is to use two RNNs in opposite directions and concatenate their representations for each word. They are used for encoding (Bahdanau et al., 2014) as well as decoding (Sennrich et al., 2016).

1.3 Limitations of MT

One important problem of current approaches in MT is that sentences in a text are treated independently, and discourse connections among them are ignored. Hardmeier (2014) presents an extensive description about this problem in SMT framework. In his work, he points out the lack of lexical cohesion, terminological inconsistency, and sometimes poor word choices. In a study made by Carpuat and Simard (2012), they analyze a English-French corpus, and show that between 15% to 25% of the phrases are inconsistent. The same issues are extended to NMT which is build over the same sentence-level paradigm. Discourse connections are also present inside a single sentence. The issues are specially evident while translating long sentences. The reason is that current MT lacks of a mechanism to model long-term dependencies among words. NMT models intra-sentence connections better than SMT, because of the recursive layer which enlarges the context of dependency, however, the problem is still present. Further analysis and examples are presented in Section 3.1.

1.4 Discourse connections for improving MT

Several approaches have being proposed in order to incorporate discourse connections in MT. Some of them identify these connections using tools such as coreference resolution. Whereas, other approaches formulate the problem as document-level MT.

1.4.1 Coreference Resolution

Coreference resolution is the task of grouping or linking mentions that refer to the same entity in a text. This task includes two stages: mention identification, and coreference resolution. The first stage is usually based on part-of-speech annotation and named-entity recognition. Candidate mentions are usually noun phrases, nouns, and pronouns (Lee et al., 2011). Coreference resolvers follow three main approaches: pairwise, re-ranking, and clustering. Pairwise resolvers perform a binary classification, predicting if two mentions refer to the same entity or not (Bengtson and Roth, 2008; Mitkov, 2002; Ng, 2010). The second approach lists a set of candidate antecedents for each mention that are simultaneously considered to find the best match (Wiseman et al., 2015). Finally, the clustering approach considers the features of a complete cluster of mentions to decide whether a mention belongs or not to that cluster (Clark and Manning, 2015; Fernandes et al., 2012). Coreference resolution is typically evaluated in comparison with a gold-standard annotation. The main metrics used for evaluation are MUC (Vilain et al., 1995), B^3 (Bagga and Baldwin, 1998), CEAF (Luo, 2005), and BLANC (Recasens and Hovy, 2011). These metrics have different methods for evaluation but they all retrieve the recall, precision, and F_1 score of the evaluated system. The performance of state-of-the-art coreference resolvers is still limited, even though, the latest progress made with neural networks is substantial (Clark and Manning, 2016; Wiseman et al., 2016).

1.4.2 Coreference-Aware Machine Translation

The interest on using coreference systems to improve translation has recently emerged (Le Nagard and Koehn, 2010; Hardmeier and Federico, 2010; Guillou, 2012). The limited accuracy of coreference resolution may explain its restricted use in MT, although, it is well known that some pronouns require knowledge of the antecedent for correct translation. For instance, Le Nagard and Koehn (2010) trained an English-French translation model where some English pronouns were manually annotated with the gender of their antecedent on the target side. The system translates well 70% of these pronouns, but it did not beat the baseline MT. Recently, a model for MT decoding proposed by Luong and Popescu-Belis (2016); Luong et al. (2015b) used in a probabilistic way several features of the antecedent candidates (e.g. gender, number and humanness values), and demonstrated some improvement on pronouns. Two shared tasks on pronoun-focused translation have been recently organized (Hardmeier et al., 2015; Guillou et al., 2016). The results show only marginal improvement of pronoun translation respect to a baseline SMT system. The systems with the best performance used deep neural networks: Luotolahti et al. (2016) and Dabre et al. (2016) summarizing the preceding and following contexts of the pronoun to predict it.

1.4.3 Document-level Machine Translation

Document-level machine translation presents a integral solution for modeling discourse-level connections. Hardmeier et al. (2013) proposed a statistical document-level decoder based on phrase-based MT. It uses a local search approach which scores the translation of an entire document at any time. The optimization is made with a hill climbing strategy. The document is represented as a sequence of sentences, and a sentence as a sequence of anchored phrases. The initialization is made with a phrase-based MT. At each step anchored phrases are modified, and the document score is recalculated.

2 State of the Research

Coreference resolution has the potential to guide the MT towards a consistent translation of entities, where entities primarily refer to noun-phases, nouns, and pronouns. During the first year of the thesis, we develop a proof-of-concept coreference aware MT (Werlen and Popescu-Belis, 2017), with the aim to verify the improvement we can get from it. In this section, we describe the motivation behind this work, and we present two different approaches for addressing the problem. Then, we define a metric to evaluate in a specific manner the translation of nouns and pronouns given that existing metrics such as BLEU are generic, and make it difficult to compare specific translations. Finally, we present the results and analysis of the systems comparing them with two different baselines.

2.1 Coreference Resolution for MT

We formulate this work as a comparison between the coreference chains found in a source text, and its translation. In both cases, the coreference links should be identical based on the principle that the information conveyed in a text must be preserved in the translation. Consequently, we explore the hypotheses that *better translations should have coreference chains that are more similar to the source*. Table 1 illustrates this criterion on an example of Spanish-to-English translation [1]. We applied automatic coreference resolver in both translations [2], and the resulting coreference chains are indicated in the table with numbers and colors. We observe that the chains in the human translation match well those in the source, but this is less the case for the automatic translation, in particular due to wrong pronoun translations.

Source	**Human Translation**	**Machine Translation**
La película narra la historia de [un joven parisiense]$_{c_1}$ que marcha a Rumanía en busca de [una cantante zíngara]$_{c_2}$, ya que [su]$_{c_1}$ fallecido padre escuchaba siempre [sus]$_{c_2}$ canciones.	The film tells the story of [a young Parisian]$_{c_1}$ who goes to Romania in search of [a gypsy singer]$_{c_2}$, as [his]$_{c_1}$ deceased father use to listen to [her]$_{c_2}$ songs.	The film tells the story of [a young Parisian]$_{c_1}$ who goes to Romania in search of [a gypsy singer]$_{c_2}$, as [his]$_{c_2}$ deceased father always listened to [his]$_{c_2}$ songs.
Pudiera considerarse un viaje fallido, porque [∅]$_{c_1}$ no encuentra [su]$_{c_1}$ objetivo, pero el azar [le]$_{c_1}$ conduce a una pequeña comunidad...	It could be considered a failed journey, because [he]$_{c_1}$ does not find [his]$_{c_1}$ objective, but the fate leads [him]$_{c_1}$ to a small community...	It could be considered [a failed trip]$_{c_3}$, because [it]$_{c_3}$ does not find [its]$_{c_3}$ objective, but the chance leads ∅ to a small community...

Table 1: Comparison of coreference chains in the Spanish source vs. English human and machine translations. The chains are numbed c_n, and the void symbol ∅ indicates null subject pronoun.

To validate this statement, we present in Table 2 the MUC, B^3 and CEAF scores of a human translation vs. two different MT systems evaluated with an automatic coreference resolver [2]. The source is a set of documents with ca. 3.5K words with gold-standard coreference annotation [1]. The BLEU score measured on the same set of documents is 49.7 for the online commercial NMT, and 43.4 for a baseline phrase-based MT (PSMT). We can observe that the coreference scores are directly correlated with the quality of translation.

Based on this idea, we implement a proof-of-concept coreference-aware MT system for Spanish-to-English translation. This pair is particularly challenging because Spanish is a drop-pronoun language, so that an MT system must not only select the correct translation of pronouns, but it must also generate English pronouns from Spanish null ones.

[1] Data from AnCora-ES corpus (Recasens and Martí, 2010)

[2] Stanford Coreference Resolution system (Manning et al., 2014)

Metric	Translation	Recall	Prec.	F1
MUC	Human	31	46	37
	Commercial NMT	21	38	28
	Baseline PSMT	18	33	23
B^3	Human	24	49	32
	Commercial NMT	20	38	26
	Baseline PSMT	17	40	24
CEAF	Human	41	40	41
	Commercial NMT	34	39	36
	Baseline PSMT	32	35	33

Table 2: Coreference similarity scores (%) between source and target texts for different translations.

2.2 Using Coreference Similarity to Rerank MT Hypotheses

We defined a document-level coreference similarity score $C_{sim}(d_t, d_s)$, as the average of MUC, B^3 and CEAF scores, to compare the coreference chains in the source document d_s and its translation d_t. We use this score to rerank the n-best translation hypotheses of each sentence coming from the MT system. The coreference similarity is not measured individually for each sentence, but at the document level. The goal is to find a combination of translations that optimizes this global score. We model problem as follows: A translated document d_t is represented as an array of translations $d_t = (s^1, s^2, ..., s^M)$, where each sentence can be choose from a list of n-best translation hypotheses $s^i \in \{s^i_1, s^i_2, ..., s^i_N\}$. The objective is to select the best combination of hypotheses based on their coreference similarity with the source, i.e.:

$$\underset{a_1,a_2,..,a_M}{\arg\max}\, C_{sim}((s^1_{a_1}, s^2_{a_2}, ..., s^m_{a_M}), d_s)$$

We try reduce the search space to allow reasonable performance. First, sentences with duplicate set of mentions are filtered out. Second, we apply beam search optimization based on the fact that the first mentions of entities usually contain more information than the next ones. The search starts from the first sentence and aggregates at each step the translation hypothesis with the highest similarity scores with the preceding ones.

There are some limitations of this approach. First, with a sentence containing several mentions, there is no guarantee that the n-best hypotheses include a combination of mention translations that optimize all mentions as the same time. Second, the correct translation of a given mention may not present at all among the n-best hypotheses, because the differences among the top hypotheses are often very small, especially when sentences are long. In order to address these issues, we present a second approach.

2.3 Post-editing Mentions Based on Cluster-level Coreference Score

This approach uses translation hypotheses of individual mentions rather than of complete sentences. This allows to optimize the translation of each mention independently, and to increase the variety of hypotheses. Additionally, instead of searching for similar clustering in the target side, we try to induce it. Thus, we define a cluster-level coreference score C_s that represents the likelihood that all mentions in that cluster refer to the same entity. We rely on the coreference resolver applied to the source side to define the clusters of mentions. Each cluster is defined as a set of mentions $c_x = \{m^i, m^j, .., m^k\}$, where each mention can be selected from a set of translation hypotheses $m^i \in \{m^i_1, m^i_2, ..., m^i_N\}$. The objective is to find the combination of translation hypotheses of mentions with higher cluster-level coreference score. We simplify the optimization with a beam search where mentions in a cluster are processed one at a time. A new upcoming mention is compared with each of the previously selected ones with a pair scorer (Clark and Manning, 2015). This scorer uses a logistic classifier to assign the probability to a pair of hypotheses representing the likelihood that they are correferent. It is defined as follows:

$$p_{pair}(m^i_{a_i}, m^j_{a_j}) = (1 + e^{\theta^T f(m^i_{a_i}, m^j_{a_j})})^{-1}$$

where $f(m^i_{a_i}, m^j_{a_j})$ is a vector of feature functions of the mentions and θ is a vector of weights. Then, C_s is defined as the product of the individual pairwise probabilities, and it is scaled with respect to all combinations. At each step, combinations with lower C_s are pruned, and the algorithm continues until precessing the last mention in the cluster.

In order to enhance the decision process, we include two sources of additional information: the translation frequency, that can help to decide between synonymous words by selecting the most frequently translation version; and the source-side entity features, which enriches the knowledge about the entity itself. The translation score T_s of a hypotesis is calculated based on its relative frequency of emission by the MT system, as follows: $T_s(m^i_{ai}) = count(m^i_{ai}) / \sum_j count(m^i_j)$. Additionally, we summirize the features of the entity in the source-side from all its mentions. Then, we define a simple scoring function which measures how well a hypotesis match with those entity's features, with respect to other alternatives: $E_s(m^i_{ai}) = f(m^i_{a_i}, \theta_{e_x}) / \sum_j f(m^i_j, \theta_{e_x})$ where f is a linear function and θ_{e_x} are the entity features. Finally, the decision is made through the weighted combination of the three previous scores:

$$C_{score}(m^i_{a_i}, m^j_{a_j}, ...) = C_s(m^i_{a_i}, m^j_{a_j}, ...)^{\lambda_1} \times [T_s(m^i_{a_i}).T_s(m^j_{a_j})...]^{\lambda_2} \times [E_s(m^i_{a_i}).E_s(m^j_{a_j})...]^{\lambda_3}$$

where $\sum_i \lambda_i = 1$. The weights λ are hyper-parameter of the function tuned according the data.

2.4 Measuring the Accuracy of Pronoun Translation

Measuring the pronoun translation is difficult, due to the interplay between the translation of pronouns and of their antecedents, and to variations in the use of non-referential pronouns. Human evaluation come at a significant cost, and its principle does not allow repeated evaluations with new candidate sentences. We propose a simple, reference-based metric to estimate the accuracy of pronoun translation (APT) (Werlen and Popescu-Belis, 2016). This metric relies on a reference human translation and on word alignment. It compares each candidate against the corresponding reference, assuming that, at least when averaged over a large number of instances, a pronoun is well translated when it is identical to the reference. Partial matches can also contribute to the score. They are defined using equivalence classes that can be learn from manual evaluation samples. The probability of a correct equivalence of different pronouns is defined as $p(c = 1|t, r)$ where t and r is the parallel pair candidate and reference pronoun respectively, $r <> t$, and c corresponds to the manual evaluation score (0 incorrect, 1 correct). The metric was applied to the results of seven systems that participated in the DiscoMT 2015 shared task on pronoun translation from English to French (Hardmeier et al., 2015). It reaches around 0.993–0.999 Pearson correlation with human judges, while other automatic metrics such as BLEU, METEOR, or those specific to pronouns used at DiscoMT 2015 reach 0.972–0.986. This metric can be adapt to other languages, as shown in the experimental part, we use it for the Spanish-English data-set. Additionally, we modify and simplify it to measure the accuracy of noun translation (ANT) based on complete matches against the reference.

2.5 Baseline Systems

We built a set of baseline systems to be used in different experiments. For the PSMT baseline, we use the Moses toolkit (Koehn et al., 2007), while for NMT we use a modification of the Deep Learning for MT tutorial system (Cho, 2016) that implements RNN encoder-decoder with attentions mechanism. We use Spanish-English training data from the translation task of the

WMT 2013 workshop (Bojar et al., 2013a). Table 3 shows the BLEU score of these two systems training on different sizes of data. The testing, *News Test 2013*, consist of approximately 3K sentences.

System	Training	Tuning	Language Model	BLEU
PSMT$_1$	1.9 M	5 K	3-gram 1.9 M	24.51
NMT$_1$	1.9 M	5 K	None	21.53
PSMT$_2$	7.6 M	5 K	3-gram 7.6 M	25.43
NMT$_2$	7.6 M	5 K	None	25.65
PSMT$_3$	14 M	5 K	4-gram 17 M	30.81
NMT$_3$	14 M	5 K	None	32.21

Table 3: Baselines PSMT and NMT systems. The data is given in number of sentences in millions (M) and thousands (K).

2.6 Experimental Results

We test the two proposed methods re-ranking and post-editing vs. the baseline PSMT$_3$. Additionally, we include the baseline NMT$_3$ as a reference for comparison only. We choose to built the two systems over a PSMT baseline for simplicity because the word-alignment can be obtained directly from the system. The word-alignment is needed for matching mentions from source and target texts, and to compare them. Tables 4 shows the results of the experiments. We first calculate BLEU, APT, and ANT values at document-level, and show the values of the average and standard deviation for our baseline, and our two proposed approaches with manual annotated coreference resolution in the source side, and the last proposed approach using instead an automatic coreference resolver. Additionally, we show the significance levels (t-test) of the results in comparison to the baseline: generally above, but one one occasion (BLEU scores of re-ranking vs. baseline) below the baseline. The post-editing approach improves the pronoun translation quite significantly, without decreasing the overall quality of translation. This improvement is demonstrated by the rise of APT scores as well as human evaluation scores (although the latter target all nouns, but in practice most of the changes are observed on pronouns). Moreover, a comparative analysis of human scores for each mention in fourth of the evaluated documents (not visible in the tables) shows that post-editing improves the translation in 45 cases but degrades it in 15 cases; the net improvement is thus 30 occurrences out of 189 examined by the human evaluator. The re-ranking approach, despite the theoretical appeal of its definition, fails to improve noun and pronoun translation. As for post-editing, it appears that the quality of the translation of nouns does not change significantly, as shown by the ANT and BLEU metrics. Finally, we see that after applying an automatic coreference resolver the quality of pronoun translation still increases, even though, we see a small but significant degradation in the BLEU score. We can attribute this to the mistakes introduced by the automatic resolver.

2.7 Conclusion

During this first year, we have explore how coreference resolution can help to improve the translation of mentions from a conceptual perspective. We have shown that with a post-editing approach we can have some improvement specially in the case of pronouns. However, the system has several restrictions. In first place, the process to obtain candidate translations requires a second pass of the text to the MT and, in some cases, the correct translations are not part of the hypotheses. Secondly, the final probabilities are a combination of independent probabilities which requires extra hyper parameters. We believe that these problems can be addressed by the integration of coreference resolution with MT. In this way, the access to hypotheses is not restricted, and the optimization can be done by taking into account all affecting probabilities at the same time while decoding.

	System				
Metric (Baseline)	**PSMT**	**NMT**	**PSMT +** **Re-rank**	**PSMT +** **Post-edit**	**PSMT +** **Post-edit** **(automatic CR)**
BLEU	46.5±4.3	46.9±3.7	41.7±3.9***	46.4±3.9	46.1±4.3
APT	0.35±0.07	0.37±0.07	0.40±0.10*	0.59±0.13***	0.41±0.07*
ANT	0.78±0.08	0.78±0.07	0.74±0.01**	0.78±0.07	0.76±0.09

Table 4: Comparison of baseline MT and our proposals for re-ranking or post-editing, for three metrics. In addition to the average scores and standard deviation over the ten test documents, we indicate the statistical significance level of the difference between each of our systems and the baseline (* for 95.0%, ** for 99.0% and *** for 99.9%).

3 Research Plan

In the following years, we aim to enhance MT by modeling discourse-level dependencies. We will direct our efforts towards NMT as it has demonstrated to increase the quality of translation compared to previous approaches, and still has potential for further improvement. In the first part of the future work, we will focus on extending the capacity of the current sentence-level NMT by incorporating long-range dependencies, and giving the network the possibility to access contextual information from previous sentences. In the second part, we intend to approach the problem of document-level MT by proposing a new architecture based on hierarchical RNNs, and the integration of a coreference resolution mechanism to keep track of entities.

3.1 Enhanced Sentence-Level NMT

Neural networks are known to learn internal representations and patterns in the data given sufficient resources. Current NMT, however, lacks of a mechanism to model long-term dependencies, and it does not have access to contextual information. By incorporating these elements, the quality of translation can potentially improve. At this stage, we do not aim to integrate coreference resolution in translation given that this task is made at document-level.

3.1.1 Long-Range Dependencies

In a sequence modeling framework such as RNN or LSTM, the input sequence is compressed into one single vector representation. Therefore, as the sentence grows, the network loses its capability to maintain all the information presented to it. This gives rise to at least two problems for modeling word connections: (1) missing long-range dependacies, and (2) undesirable sharing of attributes among words.

Current networks can capture dependencies among words in simple scenarios (e.g. having one single entity, or few entities sharing similar attributes), but they can not model dependacies in more complex scenarios. We can see an example[3] of this issue in the following translation from English-to-French[4]. In the first sentence, where all mentions are feminine, we can see that there is agreement between nouns-pronouns, nouns-adjectives, and nouns-verbs. However, this is not the case in the second sentence where there are entities with different gender. Here, we see that there is local noun-verb agreement, but, as shown in red, several long-range dependencies are missed (i.e. noun-pronoun, noun-adjective).

Source: *When she ran down,* ***the left slipper*** *remained* ***stuck, it*** *was* ***small*** *and* ***dainty****.*
MT: *Quand elle courait,* ***la pantoufle gauche*** *restait* ***collée, elle*** *était* ***petite*** *et* ***délicate****.*

[3] Taken from "Cinderella" `http://stenzel.ucdavis.edu/180/anthology/aschenputtel.html`
[4] The translation was made with a free online NMT system.

Source: *When she ran down, **the left slipper** remained **stuck**. The king's son picked **it** up, and noticed that **it** was **small** and **dainty**.*
MT: *Quand elle a couru, **la pantoufle gauche** est **restée coincée**. Le fils du roi **l'a 'ramassé**, et a remarqué qu'**il** était **petit** et **délicat**.*

Moreover, the representation of a sentence in a single vector makes it difficult for the network to characterize the connections of different words in an independent manner. As a result, some attributes are wrongly spread over words. For example in the following sentence, if we change the pronoun *"she"* by the masculine *"he"* in the source sentence, it changes the article of *"the left slipper"* making it also masculine in the translation.

Source: *When he ran down, **the left slipper** remained **stuck**, **it** was **small** and **dainty**.*
MT: *Quand il a couru, **le pantoufle gauche** restait **collé**, **il** était **petit** et **délicat**.*

In a RNN, the prediction of a word is conditioned on the previous hidden state and the input. We suggest to enrich this conditioning by including not only the previous hidden state but all hidden states of words in which it depends (Miculicich Werlen et al., 2017). The network should be able to identify which are the dependencies of the word that is being processed at a given time. This can be done by using different techniques for instance by attention mechanism or by keeping internal memory representations. The aim is to increase the contextual information available to make a better prediction.

The first approach that we will explore is inspired by the work of Cheng et al. (2016) who proposed an enhanced LSTM for language modeling, with the goal of better capturing linguistic structures in sentences. We plan to use a similar technique to capture word dependencies. When predicting a new word, the conditioning probability of the hidden state will be enriched by an attention mechanism over the hidden states of the previously predicted words. Encouraged by the positive results presented by Cheng et al. (2016), we expect that the mechanism will learn to discriminate useful information from the past for a particular word, and capture the linguistic dependencies such as noun-pronoun and noun-adjective agreements, repeated words, use of synonyms, etc. Figure 3 shows the baseline architecture for NMT. Figure 4 shows the proposed architecture which incorporates an additional attention mechanism in the decoder as described earlier.

More advanced approaches will also be explored. The identification of long-range dependencies requires understanding of the morphology, syntax, and semantics. The RNN architecture is able to detect simple patterns in sequences, but it does not capture linguistic structure to such a level of complexity. Thus, we plan to enhance it with the use of an external memory. This has been researched in different architectures (Weston et al., 2014; Sukhbaatar et al., 2015; Rae et al., 2016; Gulcehre et al., 2017; Cheng et al., 2016), and has empirically shown improvements over simple RNNs for different tasks such as language modeling and question answering. It gives the network the capacity to process complex tasks that involve long-term dependencies, and reasoning over text. One related application is reading comprehension, as proposed by Hermann et al. (2015), where the aim is to guess a hidden noun in a document by using only contextual information.

3.1.2 Incorporating Context

The current restriction of NMT to the sentence-level is due to the difficulty of RNNs to manage long sequences. The back-propagation algorithm used for learning unfolds the network over the sequence, and the vanishing gradient problem increases with the length of the sequence. We present here an example[5] of the problem of translating texts as independent sentences. It is a Spanish-to-English translation[6] the Spanish word *"partido"* has two senses, which can be translated into different words in English. In the second line of the dialogue, the MT system translates

[5] Example taken from Opus-Subtitles-2016 `http://opus.lingfil.uu.se/OpenSubtitles2016.php`

[6] The translation was made with a free online NMT system.

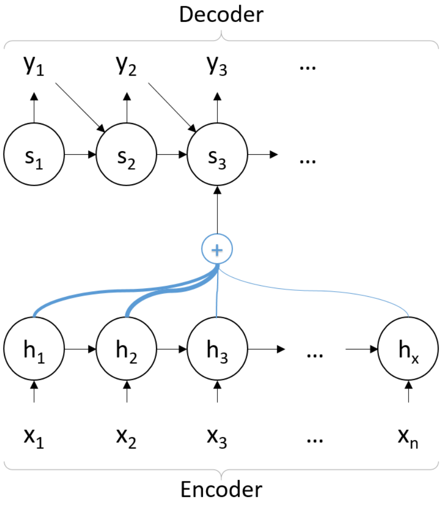


Figure 3: Baseline NMT: Encoder-decoder RNN architecture with attention mechanism.

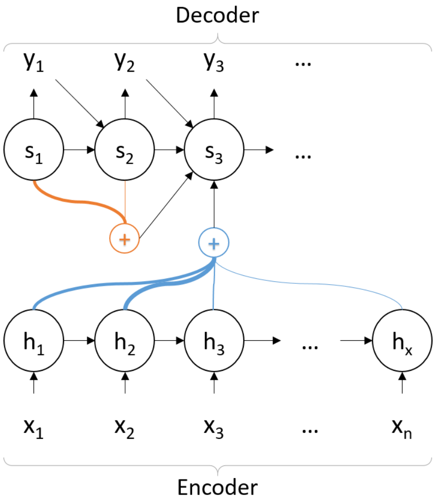


Figure 4: First approach for including long-range dependencies. NMT + attention mechanism over previous words.

it incorrectly because the context is lost.

Source:	*Pertenezco a un **partido** político respetable. – ¿Qué **partido**?*
Reference:	*I belong to a respectable political **party**. – Which **party**?*
MT:	*I belong to a respectable political **party**. – What a **match**?*

To deal with this problem, we propose to enhance the sentence-level NMT by adding contextual information coming from previous sentences. This approach is inspired by the notion of human reading process. While reading, a person forms certain concepts about the text, and use them for understanding the flow of the story. We will investigate the use of "memory" networks to capture meaningful representations of previous sentences, which are given as input to the network. The memory consists of an array of cells containing representations of informative units, for instance words and sentences. They can be written, read, and forgotten. The past sentences can be input using bag-of-words, or using the hidden representations obtained when they were translated. The challenge is to capture meaningful representations. One way is by utilizing the memory network proposed by Sukhbaatar et al. (2015) for answering questions, and language model. They obtain continuous representation of several sentences via recurrent attention mechanisms. We will use a simplified version of it. Later, while decoding, the conditioning over the hidden state of a word will be enriched with dependencies over the memory. The idea is to give the network the capacity to represent and summarize past information which is useful for current predictions. Figure 5 shows the proposed architecture.

3.2 Document-level NMT

In the previous section, we presented our plans to enhance sentence-level NMT with contextual information. In this section, we propose a more integrated solution to model discourse-level connections by formulating the problem of document-level NMT. The aim is to optimize the translation of complete documents. Document-level MT has being proposed for the phrase-based SMT approach (Hardmeier et al., 2013), but it has not been explored yet in NMT framework.

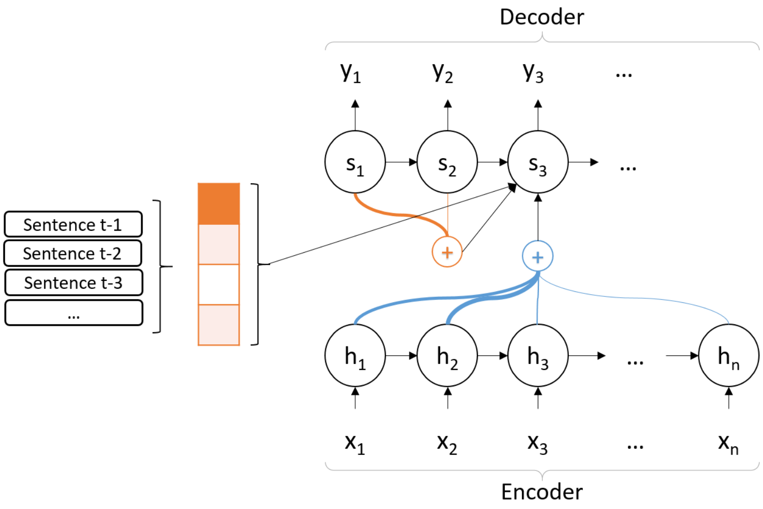


Figure 5: Including contextual information. NMT + attention mechanism over previous sentences.

3.2.1 Document Modeling with Hierarchical RNNs

We propose to use hierarchical RNNs, which consist of a number of sub-networks, arranged in layers. Each sub-network is intended to capture a particular aspect of the input data, but the patterns and relationships are determined by training the network as a whole (Ruiz and Srinivasan, 2002). In this case, each sub-network will model one sentence of the document. The higher layers could correspond to paragraphs and document representations. We intend to investigate different configurations to integrate within the existing encoder-decoder framework of NMT. Figure 6 shows a simple architecture of document-level NMT. Here, the encoder has two hierarchical levels: words and sentences; while the decoder has only one. The contextual information is modeled by the second hierarchical layer of the encoder. This layer transmit the information from one sentence to another. The decoder is similar to the one in the sentence-level NTM, but it is conditioned over the sentence vector representation in order to obtain information from previous sentences.

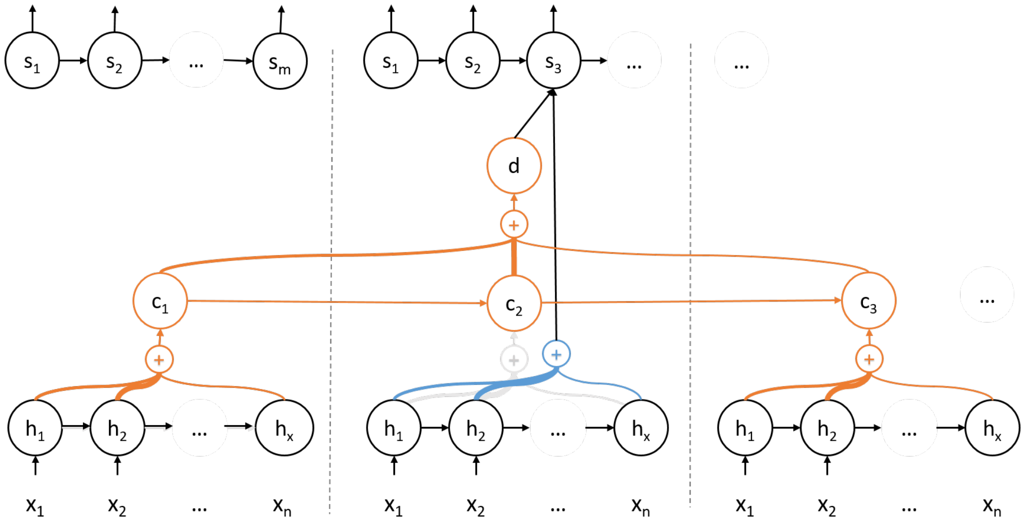


Figure 6: First proposed architecture for Document-Level NMT.

3.2.2 Coreference-aware NMT

In the last part of this thesis, we consider the integration of document-level NMT with a coreference resolution mechanism. The recognition of coreferent words is a complicated task because it involves processing of different linguistic elements such as morphology, syntax, and semantics, and even requires world knowledge in some cases. It is thus possible that even a rich document-level NMT model does not fully capture these connections.

Current automatic coreference resolvers divide the task of recognizing different linguistic elements in several successive stages: part-of-speech recognition, entity name recognition, syntactic parsing, etc. The extracted features are then passed to the coreference resolver. In the context of NMT, the features are simple words, therefore, the identification of coreferences in this condition is more demanding. However, it is not necessary to identify all coreferences, and this task should be conditioned according to the need to disambiguate the translation of a word only.

There are two sides where we can identify coreferences: source and target. Our previous strategy of coreference-aware MT was to apply coreference resolution on the source side, and use these relations to produce better translations. The reason is that the source contains complete information for the direct use of coreference resolver. Therefore, the first approach that we will explore is a multitask learning process to enrich the feature representation during encoding. The network will simultaneously optimize the translation, and the coreference resolution in the source side. The coreference resolution task can be addressed as proposed by Wiseman et al. (2016), who use RNNs to learn internal representations of entities from its mentions. One important issue is that there are no data-sets for learning both tasks at the same time. Thus, we will investigate possible solutions such as alternation of samples from the different corpus of each task, or creating synthetic-data as proposed by Sennrich et al. (2016), where they use samples of monolingual corpus, translated with an external MT, to train their NMT.

3.3 Implementation

In this section we describe the data-sets for training and testing the systems, the metrics for evaluation, and the infrastructure that will be used to implement our proposals.

3.3.1 Datasets

The various MT architectures will be tested for English-French and Spanish-English. These languages are selected for the well-known challenges for translating anaphoric pronouns. The translation of English-French is ambiguous due to the different assignation of gender and number to object. Whereas, the translation Spanish-English is challenging because Spanish is a pro-drop language, in which subject pronouns may be missing. Data for these languages is vastly available, and the quality of sentence-level translation is relatively good, which lets us focus on the problem of discourse-level constraints. The datasets we will use come from the Workshop in Machine Translation (WMT) which combines: Europarl, News Commentaries, UN and common crawl corpora. The English-French training data consists of a subset of WMT 2014 data (Bojar et al., 2014), with selected data optimized for news translation (Axelrod et al., 2011)[7]. The Spanish-English data is taken from WMT 2013 (Bojar et al., 2013b). Additionally, we plan to use a corpus of copyright free books from Tiedemann (2012). The translation of books is a challenge because it requires higher contextual information, so this type of data is useful for testing discourse connections. The data will be preprocessed, e.g. with the tokenizer and truecaser from Moses toolkit (Koehn et al., 2007). Table 5 shows the training, development, and testing sets.

[7] Sub-set available at http://www-lium.univ-lemans.fr/~schwenk/cslm_joint_paper/

Data pair	Training	Development	Testing
English-French	WMT 2014 (14 M)	news test 2012-2013 (6 K)	news test 2014 (3 k)
	Opus: 29 books (0.1 M)	TBD	TBD
Spanish-English	WMT 2013 (14 M)	news test 2010-2011 (5 K)	news test 2013 (3 k)
	Opus: 18 books (93 K)	TBD	TBD

Table 5: Datasets that will be used in this thesis. Numbers of sentences are given in millions (M) or thousands (K).

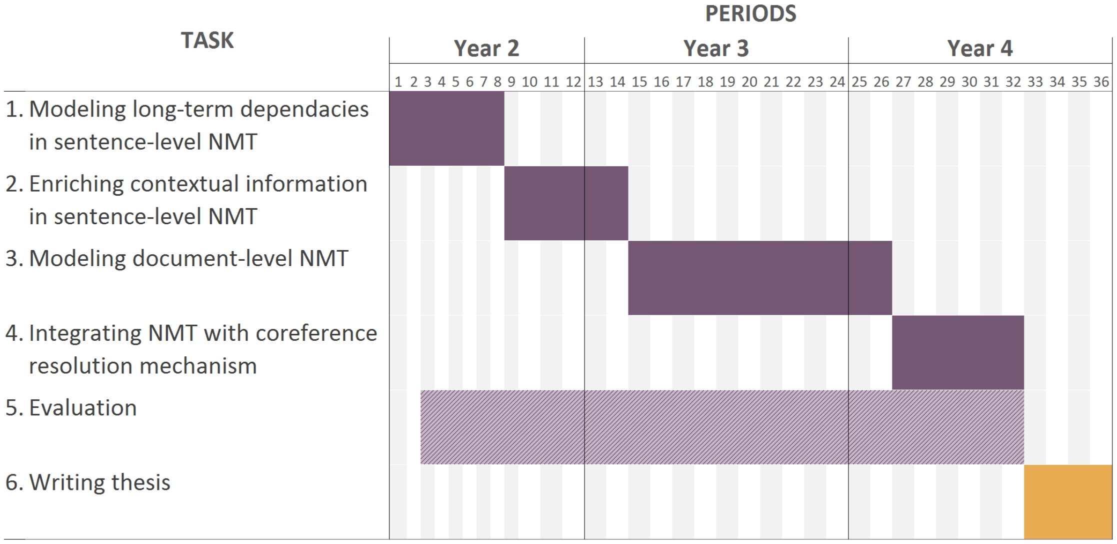
3.3.2 Evaluation Metrics

The evaluation will be done with the BLEU score (Papineni et al., 2002) with the implementation *mteval-v13a* from the Moses toolkit (Koehn et al., 2007). For particular evaluation of entity connections, we will also use Meteor (Lavie and Denkowski, 2009) with a filter for nouns and pronouns. We will also use our proposed metrics APT and ANT (see Section 2.4) as well as metrics for document-level MT Wong and Kit (2012); Xiao et al. (2011); Giménez et al. (2010).

3.3.3 Infrastructure

The training of deep neural networks like RNNs requires high performance hardware and software infrastructure. Most of the current implementations are based on graphics processing units (GPUs) (Goodfellow et al., 2016). We have a disposition several GPUs with NVIDIA's CUDA[8] a parallel computing platform and programming model for increasing computing performance. We use Theano[9] library for the implementations which is based on Python. It is a very extensible frameworks for the implementation of new architectures, and it has a very good permanence on GPU training for LSTMs compared to other languages Bahrampour et al. (2015).

4 Schedule



[8] http://www.nvidia.com/object/cuda_home_new.html
[9] http://deeplearning.net/software/theano/index.html

Publications

1. Miculicich Werlen, L. and Popescu-Belis, A. (2017). Using Coreference Links to Improve Spanish-to-English Machine Translation. In *Proceedings of the Workshop on Coreference Resolution Beyond OntoNotes (CORBON)*. Association for Computational Linguistics.
2. Luong, N. and Werlen Miculicich, L. and Popescu-Belis, A. (2015). Pronoun Translation and Prediction with or without Coreference Links. In *Proceedings of the Second Workshop on Discourse in Machine Translation*. Association for Computational Linguistics.

To be submitted to the workshop on Discourse in Machine Translation 2017:

3. Miculicich Werlen, L. and Popescu-Belis, A. (2016). Validation of an Automatic Metric for the Accuracy of Pronoun Translation (APT). Technical report. Idiap.

References

Axelrod, A., He, X., and Gao, J. (2011). Domain adaptation via pseudo in-domain data selection. In *Proceedings of the conference on empirical methods in natural language processing*, pages 355–362. Association for Computational Linguistics.

Bagga, A. and Baldwin, B. (1998). Algorithms for scoring coreference chains. In *Proceedings of the First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference*, volume 1, pages 563–566, Granada, Spain.

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

Bahrampour, S., Ramakrishnan, N., Schott, L., and Shah, M. (2015). Comparative study of deep learning software frameworks. *arXiv preprint arXiv:1511.06435*.

Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. (2003). A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155.

Bengtson, E. and Roth, D. (2008). Understanding the value of features for coreference resolution. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 294–303, Honolulu, Hawaii.

Bentivogli, L., Bisazza, A., Cettolo, M., and Federico, M. (2016). Neural versus phrase-based machine translation quality: a case study. *arXiv preprint arXiv:1608.04631*.

Bojar, O., Buck, C., Callison-Burch, C., Federmann, C., Haddow, B., Koehn, P., Monz, C., Post, M., Soricut, R., and Specia, L. (2013a). Findings of the 2013 workshop on statistical machine translation. In *Proceedings of the 2013 Workshop on Machine Translation (WMT)*, Sofia, Bulgaria.

Bojar, O., Buck, C., Callison-Burch, C., Federmann, C., Haddow, B., Koehn, P., Monz, C., Post, M., Soricut, R., and Specia, L. (2013b). Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 1–44, Sofia, Bulgaria. Association for Computational Linguistics.

Bojar, O., Buck, C., Federmann, C., Haddow, B., Koehn, P., Leveling, J., Monz, C., Pecina, P., Post, M., Saint-Amand, H., Soricut, R., Specia, L., and Tamchyna, A. (2014). Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA. Association for Computational Linguistics.

Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., Logacheva, V., Monz, C., et al. (2016). Findings of the 2016 conference on machine translation (wmt16). In *Proceedings of the First Conference on Machine Translation (WMT)*, volume 2, pages 131–198.

Brown, P. F., Cocke, J., Pietra, S. A. D., Pietra, V. J. D., Jelinek, F., Lafferty, J. D., Mercer, R. L., and Roossin, P. S. (1990). A statistical approach to machine translation. *Computational linguistics*, 16(2):79–85.

Carpuat, M. and Simard, M. (2012). The trouble with smt consistency. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 442–449. Association for Computational Linguistics.

Cheng, J., Dong, L., and Lapata, M. (2016). Long short-term memory-networks for machine reading. *arXiv preprint arXiv:1601.06733*.

Chiang, D. (2005). A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 263–270. Association for Computational Linguistics.

Cho, K. (2015). Introduction to neural machine translation with GPUs (part 3). `https://devblogs.nvidia.com/parallelforall/introduction-neural-machine-translation-gpus-part-3/`.

Cho, K. (2016). dl4mt-tutorial. `https://github.com/nyu-dl/dl4mt-tutorial`.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Clark, K. and Manning, C. D. (2015). Entity-centric coreference resolution with model stacking. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1405–1415, Beijing, China.

Clark, K. and Manning, C. D. (2016). Improving coreference resolution by learning entity-level distributed representations. *arXiv preprint arXiv:1606.01323*.

Dabre, R., Puzikov, Y., Cromieres, F., and Kurohashi, S. (2016). The Kyoto University cross-lingual pronoun translation system. In *Proceedings of the First Conference on Machine Translation*, pages 571–575, Berlin, Germany.

Devlin, J., Zbib, R., Huang, Z., Lamar, T., Schwartz, R. M., and Makhoul, J. (2014). Fast and robust neural network joint models for statistical machine translation. In *ACL (1)*, pages 1370–1380. Citeseer.

Fernandes, E. R., dos Santos, C. N., and Milidiú, R. L. (2012). Latent structure perceptron with feature induction for unrestricted coreference resolution. In *Proceedings of the Joint Conference on EMNLP and CoNLL - Shared Task*, pages 41–48, Stroudsburg, PA, USA.

Giménez, J., Màrquez, L., Comelles, E., Castellón, I., and Arranz, V. (2010). Document-level automatic mt evaluation based on discourse representations. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 333–338. Association for Computational Linguistics.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`.

Guillou, L. (2012). Improving pronoun translation for statistical machine translation. In *Proceedings of EACL 2012 Student Research Workshop (13th Conference of the European Chapter of the ACL)*, pages 1–10, Avignon, France.

Guillou, L., Hardmeier, C., Nakov, P., Stymne, S., Tiedemann, J., Versley, Y., Cettolo, M., Webber, B., and Popescu-Belis, A. (2016). Findings of the 2016 WMT Shared Task on cross-lingual pronoun prediction. In *Proceedings of the First Conference on Machine Translation*, pages 525–542, Berlin, Germany.

Gulcehre, C., Chandar, S., and Bengio, Y. (2017). Memory augmented neural networks with wormhole connections. *arXiv preprint arXiv:1701.08718*.

Hardmeier, C. (2014). *Discourse in statistical machine translation*. PhD thesis, Acta Universitatis Upsaliensis.

Hardmeier, C. and Federico, M. (2010). Modelling pronominal anaphora in statistical machine translation. In *Proceedings of International Workshop on Spoken Language Translation (IWSLT)*, Paris, France.

Hardmeier, C., Nakov, P., Stymne, S., Tiedemann, J., Versley, Y., and Cettolo, M. (2015). Pronoun-focused MT and cross-lingual pronoun prediction: Findings of the 2015 DiscoMT shared task on pronoun translation. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 1–16, Lisbon, Portugal.

Hardmeier, C., Stymne, S., Tiedemann, J., and Nivre, J. (2013). Docent: A document-level decoder for phrase-based statistical machine translation. In *ACL 2013 (51st Annual Meeting of the Association for Computational Linguistics); 4-9 August 2013; Sofia, Bulgaria*, pages 193–198. Association for Computational Linguistics.

Hermann, K. M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., and Blunsom, P. (2015). Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, pages 1693–1701.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Jelinek, F. (1980). Interpolated estimation of markov source parameters from sparse data. In *Proc. Workshop on Pattern Recognition in Practice, 1980.*

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 177–180, Prague, Czech Republic.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 48–54, Edmonton, Canada.

Lavie, A. and Denkowski, M. J. (2009). The meteor metric for automatic evaluation of machine translation. *Machine translation*, 23(2-3):105–115.

Le Nagard, R. and Koehn, P. (2010). Aiding pronoun translation with co-reference resolution. In *Proceedings of the Joint 5th Workshop on Statistical Machine Translation and Metrics (MATR)*, pages 258–267, Uppsala, Sweden.

Lee, H., Peirsman, Y., Chang, A., Chambers, N., Surdeanu, M., and Jurafsky, D. (2011). Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 Shared Task. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 28–34, Stroudsburg, PA, USA.

Luo, X. (2005). On coreference resolution performance metrics. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 25–32, Vancouver, B.C., Canada.

Luong, M.-T., Pham, H., and Manning, C. D. (2015a). Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025.*

Luong, N. Q. and Popescu-Belis, A. (2016). Improving pronoun translation by modeling coreference uncertainty. In *Proceedings of the First Conference on Machine Translation*, pages 12–20, Berlin, Germany.

Luong, N. Q., Werlen, L. M., and Popescu-Belis, A. (2015b). Pronoun translation and prediction with or without coreference links. *DISCOURSE IN MACHINE TRANSLATION*, page 94.

Luotolahti, J., Kanerva, J., and Ginter, F. (2016). Cross-lingual pronoun prediction with deep recurrent neural networks. In *Proceedings of the First Conference on Machine Translation*, pages 596–601, Berlin, Germany.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, MD, USA.

Miculicich Werlen, L., Pappas, N., Ram, D., and Popescu-Belis, A. (2017). Global-context neural machine translation through target-side attentive residual connections. Idiap-RR Idiap-RR-24-2017, Idiap.

Mitkov, R. (2002). *Anaphora Resolution.* Longman, London, UK.

Ng, V. (2010). Supervised noun phrase coreference research: The first fifteen years. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1396–1411, Uppsala, Sweden.

Olah, C. (2015). Understanding lstm networks. `http://colah.github.io/posts/2015-08-Understanding-LSTMs/`.

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 Annual Meeting on Association for Computational Linguistics (ACL)*, pages 311–318, Philadelphia, PA, USA.

Pascanu, R., Mikolov, T., and Bengio, Y. (2013). On the difficulty of training recurrent neural networks. *ICML (3)*, 28:1310–1318.

Pouget-Abadie, J., Bahdanau, D., van Merriënboer, B., Cho, K., and Bengio, Y. (2014). Overcoming the curse of sentence length for neural machine translation using automatic segmentation. *arXiv preprint arXiv:1409.1257.*

Rae, J., Hunt, J. J., Danihelka, I., Harley, T., Senior, A. W., Wayne, G., Graves, A., and Lillicrap, T. (2016). Scaling memory-augmented neural networks with sparse reads and writes. In *Advances In Neural Information Processing Systems*, pages 3621–3629.

Recasens, M. and Hovy, E. (2011). Blanc: Implementing the rand index for coreference evaluation. *Natural Language Engineering*, 17(04):485–510.

Recasens, M. and Martí, M. A. (2010). Ancora-co: Coreferentially annotated corpora for spanish and catalan. *Language Resources and Evaluation*, 44(4):315–345.

Ruiz, M. E. and Srinivasan, P. (2002). Hierarchical text categorization using neural networks. *Information Retrieval*, 5(1):87–118.

Schwenk, H., Dchelotte, D., and Gauvain, J.-L. (2006). Continuous space language models for statistical machine translation. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 723–730.

Sennrich, R., Haddow, B., and Birch, A. (2015). Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Sennrich, R., Haddow, B., and Birch, A. (2016). Edinburgh neural machine translation systems for wmt 16. *arXiv preprint arXiv:1606.02891*.

Sukhbaatar, S., Weston, J., Fergus, R., et al. (2015). End-to-end memory networks. In *Advances in neural information processing systems*, pages 2440–2448.

Tiedemann, J. (2012). Parallel data, tools and interfaces in opus. In *LREC*, volume 2012, pages 2214–2218.

Vilain, M., Burger, J., Aberdeen, J., Connolly, D., and Hirschman, L. (1995). A model-theoretic coreference scoring scheme. In *Proceedings of the 6th Conference on Message Understanding*, pages 45–52, Columbia, MD, USA.

Werlen, L. M. and Popescu-Belis, A. (2016). Validation of an automatic metric for the accuracy of pronoun translation (APT). Technical report, Idiap.

Werlen, L. M. and Popescu-Belis, A. (2017). Using coreference links to improve spanish-to-english machine translation. In *Proceedings of the Workshop on Coreference Resolution Beyond OntoNotes (CORBON)*, number EPFL-CONF-225953. Association for Computational Linguistics.

Weston, J., Chopra, S., and Bordes, A. (2014). Memory networks. *arXiv preprint arXiv:1410.3916*.

Wiseman, S., Rush, A. M., Shieber, S., and Weston, J. (2015). Learning anaphoricity and antecedent ranking features for coreference resolution. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1416–1426, Beijing, China.

Wiseman, S., Rush, A. M., and Shieber, S. M. (2016). Learning global features for coreference resolution. *arXiv preprint arXiv:1604.03035*.

Wong, B. and Kit, C. (2012). Extending machine translation evaluation metrics with lexical cohesion to document level. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1060–1068. Association for Computational Linguistics.

Xiao, T., Zhu, J., Yao, S., and Zhang, H. (2011). Document-level consistency verification in machine translation. In *Machine Translation Summit*, volume 13, pages 131–138.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R. S., and Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. *arXiv preprint arXiv:1502.03044*, 2(3):5.