Detecting Document-level Context Triggers to Resolve Translation Ambiguity

Laura Mascarell, Mark Fishel and **Martin Volk**
Institute of Computational Linguistics
University of Zurich
Switzerland
{mascarell,fishel,volk}@cl.uzh.ch

Abstract

Most current machine translation systems translate each sentence independently, ignoring the context from previous sentences. This discourse unawareness can lead to incorrect translation of words or phrases that are ambiguous in the sentence. For example, the German term *Typen* in the phrase *diese Typen* can be translated either into English *types* or *guys*. However, knowing that it co-refers to the compound *Körpertypen* ("body types") in the previous sentence helps to disambiguate the term and translate it into *types*. We propose a method of automatically detecting document-level trigger words (like *Körpertypen*), whose presence helps to disambiguate translations of ambiguous terms. In this preliminary study we analyze the method and its limitations, and outline future work directions.

1 Introduction

Words with ambiguous senses and translations pose a core challenge for machine translation. For example, the English noun *face* is translated into German *Gesicht* ("front of head") or *Wand* ("wall") when talking about mountaineering. Phrase-based Statistical Machine Translation (SMT) systems benefit from using the local context inside the phrases for disambiguation; on the other hand, global sentence-level and document-level context remains largely unmodelled. We focus on cases where the source of disambiguation lies in the sentences preceding the ambiguous term, for example:

> ...on the unclimbed ***East face*** of the Central Tower...
> ...we were swept from the ***face*** by a five-day storm...

Mascarell et al. (2014) and Pu et al. (2015) tackle the issue illustrated in the previous example, and show improvements in correctness, based on the one-translation-per-discourse hypothesis (Carpuat, 2009). Specifically, their method uses the translation of the head of the compound (e.g. *Wand* in *East face*) for the term (e.g. *face*) that co-refers back to it in a later sentence.

Bridging Noun Phrases (NPs) are a similar phenomenon that crosses sentence boundaries:

> The company wrote out ***a new job***.
> ***Two applicants*** were suitable.

Here the bridging NP *two applicants* is ambiguous on its own, as *applicants* can be translated into Spanish as *candidatos* or *solicitantes*. However, in the context of the antecedent of the bridging NP, *a new job*, *applicants* is more appropriately translated into *candidatos*.

In this work we generalize over both these problems (i.e. co-referent compounds and bridging NPs) and disambiguate translations using "trigger words": words whose presence in the preceding sentences indicates a certain context for the ambiguous term in the current sentence. We focus on automatically detecting such trigger words universally without focusing on a single phenomenon like compound co-references or bridging, and analyze the results.

2 Detecting Context Triggers

Ambiguous words with several possible translations have a different translation distribution depending on the sense; for example, the English *driver* in the meaning of the person driving a vehicle will likely be translated into the French *conducteur* or *chauffeur*, and much less likely into *pilote*, which corresponds to the computer device-related meaning. However, when estimated on the whole corpus the likelihoods of the three transla-

Proceedings of the Second Workshop on Discourse in Machine Translation (DiscoMT), pages 47–51,
Lisbon, Portugal, 17 September 2015.

German word:	BILD		LAND		TYP		FLÄCHE	
Translation distributions in different documents:	**doc. #1:**		**doc. #4:**		**doc. #7:**		**doc. #10:**	
	picture:	0.93	country:	0.84	guy:	0.33	surface:	0.93
	frame:	0.04	state:	0.09	jimbo:	0.33	faces:	0.07
	understanding	0.03	arab:	0.07	person:	0.33		
	doc. #2:		**doc. #5:**		**doc. #8:**		**doc. #11:**	
	image:	1.00	country:	1.00	type:	1.00	area:	1.00
	doc. #3:		**doc. #6:**		**doc. #9:**		**doc. #12:**	
	image:	0.73	country:	0.94	guy:	1.00	area:	0.80
	imagery:	0.22	nation:	0.03			space:	0.20
	picture:	0.05	desolate:	0.03				

Table 1: The four ambiguous words selected for our experiments from the WIT3 corpus. The table shows how the translation distribution of each word differ from document to document. Some translations are noise due to wrong word alignments.

tions in $P(\cdot|driver)$ will reflect the frequency of usage and not the particular contexts.

We focus on trigger words that appear in the context of a particular word sense. Identifying them helps to disambiguate the sense of an ambiguous word and translate it correctly. We try to detect trigger words from the preceding context and use them as conditional variables in the translation distributions. This means, for example, that

$$p(tgt = \text{“pilote”}|src = \text{“driver”}, trig = \text{“road”})$$

should be low, while

$$p(tgt = \text{“pilote”}|src = \text{“driver”}, trig = \text{“device”})$$

should be much higher (where src is the source word, tgt – its translation hypothesis and $trig$ – the trigger word).

To identify those trigger words we consider a simplistic method based on translation distribution similarity. The core idea is that the translation distribution of an ambiguous word changes with the presence and absence of a trigger word. That is, non-trigger words (e.g. function words and general vocabulary) lead to similar distributions (i.e. their presence and absence has little effect on the translation choice), whereas relevant triggers result in these two distributions being highly different. To measure this distribution difference we compute the KL-divergence between them. In other words, for each ambiguous term A we are searching for such a trigger word W from the preceding sentences that maximizes

$$D_{KL}(P(\cdot \mid A, W) \mid\mid P(\cdot \mid A, -W)),$$

where $-W$ means the absence of the trigger word W from the preceding sentences.

3 Experiments

In this preliminary evaluation of our method we focus on the specific case of co-references to compounds, where the co-reference is an ambiguous word with several translations. The co-reference is disambiguated using a trigger word from the preceding context (i.e. the compound that the word co-refers to). The idea is that knowing which these compounds are we assess whether our method is able to detect them as relevant triggers.

The data comes from the German-English part of the WIT3 corpus (Cettolo et al., 2012), which is a collection of TED talks in multiple languages. The corpus consists of 194'533 sentences and 3.6 million tokens split into 1'596 talks (i.e. documents). The test set is also a collection of TED talks, consisting of 6'047 sentences and about 100'000 tokens. The talks differ greatly in terms of the covered topics, and therefore, have a high potential for ambiguous translations between them. This topic variety is so high that it is not feasible to tune SMT systems separately to each topic. However, it makes the corpus a feasible target for dynamic adaptation like our method.

For our experiments we first manually select four ambiguous words, and we then obtain the co-referenced compounds by applying the detection method described in (Mascarell et al., 2014). Next, we check whether our method detects these compounds as triggers. The four selected words *Bild*, *Land*, *Typ* and *Fläche* are presented in Table 1; as the table shows their translation distributions indeed differ between different documents.

TOP HIGHEST DISTANCE				TOP LOWEST DISTANCE			
Lemma	EN	Score	Freq.	Lemma	EN	Score	Freq.
unterseeboot	submarine	28.8843	1/2	weil	because	0.0053	474/4'231
alvin	alvin	28.8843	3/5	eine	a	0.0061	6257/62'088
gap	gap	28.8843	1/7	leute	people	0.0078	485/4'543
unaufgefordert	unsolicited	28.8843	1/2	"	"	0.0222	1295/15'395

Table 2: Comparison of the lemmas with the highest and lowest KL divergence score in the context of *Land* considering the 4 preceding sentences. The *Freq.* column shows the total number of times the lemma appears in the context of *Land* over the total occurrences of that lemma in the corpus.

COMPOUND	1 SENT.		2 SENT.		3 SENT.		4 SENT.	
	pos.	△	pos.	△	pos.	△	pos.	△
Geburtsland	52'133	28.32	53'233	28.32	54'123	28.32	1'430	3.50
Lesterland	19'689	28.32	711	3.50	923	3.50	823	3.50
Entwicklungsland	4'811	24.96	6'744	24.83	8'300	24.93	9'717	24.96
Heimatland	5'483	25.30	94'095	28.33	10'358	28.30	94'084	28.40
Niemandsland	39'698	28.32	854	3.50	1'099	3.50	1'312	3.50

Table 3: Comparison of the resulting KL divergence ranking obtained considering the context of the previous sentences up to 4. The table shows the ranking position of the compounds co-referenced by *Land* in the corpus, and the difference between their distance score and the word with the highest distance.

4 Results and Analysis

We assess whether our method detects as triggers the compounds that the selected words (see Table 1) co-refer to. We do not try to detect the compounds directly because we aim at generalizing and applying our method to other phenomena, such as bridging. Since all selected words have a similar outcome, we focus on the results of *Land*.

We first analyze which are the detected triggers by our method for the word *Land*, considering the 4 previous sentences (see Table 2). Note that to detect the triggers, our method computes the distance between the translation distribution of the word *Land* when the trigger candidate appears in the context and when it does not. Therefore, the words with the highest distance score are the relevant triggers, while the words with the lowest are mostly frequent non-content words that do not give any information of the correct translation of *Land*. We also observe that none of the compounds are in the list of trigger words, but other non-related words. The reason is that these occur together with the ambiguous word (*Land*) only once, causing the distribution to contain only one translation with 1.0 probability. This distribution is then very different from the one without that infrequent faux-trigger over the rest of the document, which includes several translation variants.

The position of the compounds in the resulting KL divergence ranking is shown in Table 3, considering the context of the previous sentences up to 4. Table 3 also shows the difference between the compound score and the word with the highest distance (i.e. most relevant trigger detected).

To get a better overview, Figure 1 illustrates where the listed compounds (see Table 3) are positioned over the whole ranking in the context of the previous sentences up to 4. We observe that some of these compounds appear in the first quartile of the ranking. However, there are compounds in the bottom half of the graph, that is they are not detected as relevant trigger words.

5 Outlook for future research extensions

We observe in the analysis (see section 4) that our method is sensitive to detect non-related infrequent words as potential triggers. To solve this problem, we want to steer the search to semantically related words, instead of only filtering out infrequent words. The reason is that trigger words that only appear in the context of an ambiguous term would be detected as infrequent, and therefore, incorrectly discarded. We are then planning to combine the distribution difference (measured with the KL divergence or other metrics) with a measure of similarity between the trigger candidate and the ambiguous word. Their simi-

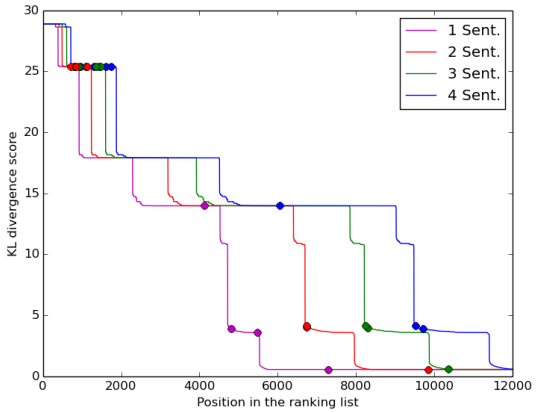

Figure 1: Comparison of the KL divergence rankings considering up to 4 previous sentences. The position of the compounds listed in table 3 are pointed out among all trigger candidates.

larity can be measured using a vector representation (Mikolov et al., 2013), for example with the $word2vec$ tool[1].

Since our method suffers from data sparsity, only trigger words that appear in the training data are taken into account. Using $word2vec$ we can compare the vector representation of the detected trigger words and the trigger candidates in the test set. We would then also consider trigger words that do not appear in the training data, but have the same vector representation.

Finally, the goal of our method is to generalize the detection of trigger words. Thus, we want to extend our study testing whether our method detects the antecedent of bridging NPs as a trigger word, and other discourse-oriented phenomena.

6 Related Work

Several approaches focus on improving lexical choice in SMT by enforcing consistency at document level. These are based on the one-translation-per-discourse hypothesis (Carpuat, 2009), which shows that more than one translation of the same term in the document leads to incorrect translations. Mascarell et al. (2014) and Pu et al. (2015) take advantage of compounds, which have more context than single-root words, and use the translation of the head of the compound for later occurrences of the single co-referring head noun in isolation. Using an enforcing and post-editing method, they show improvement of translation correctness of co-referring terms in German-French and Chinese-English. Other approaches (see (Tiedemann, 2010) and (Gong et al., 2011)) use a cache-model for the same purpose. Xiao et al. (2011) enforce the translation of ambiguous words to be consistent across the document by applying a three-steps procedure.

The term "trigger" is first introduced by Rosenfeld (1994). The approach to adaptive language modeling uses a maximum entropy model, showing perplexity improvements over the conventional trigram model.

A recently popular approach is to include topic modeling into the SMT pipeline and to use topic distributions to disambiguate phrase translations (see e.g. (Hasler et al., 2014)). Xiong et al. (2014) present a sense-based translation model that integrates word senses using maximum entropy classifiers. Meng et al. (2014) propose three term translation models to disambiguate, enforce consistency and guarantee integrity. Finally, Xiong et al. (2013) introduce a method that translates the coherence chain of the source, and uses it to produce a coherent translation. This topic modeling line of research can be combined with our own by including preceding sentences or their parts into the topic model training process.

7 Conclusions

We present a method that crosses sentence boundaries to automatically detect the words that help to correctly translate terms with several senses. We call them trigger words, and they appear in the context of a particular word sense. To detect them we compute the distance between the translation distributions of the ambiguous word with and without the presence of the trigger candidate. Higher distances suggest a likely trigger for a particular word sense.

There are two main issues that need to be solved. First, infrequent non-related trigger candidates that appear in the context of the word obtain a high distance score, and therefore, they are detected as potential triggers. Second, only the triggers detected in the training data can be used in the test set. To solve these issues, we are planning to use word vector representations to include the measurement of semantic relatedness between the ambiguous word and its triggers.

[1] https://code.google.com/p/word2vec/

References

Marine Carpuat. 2009. One translation per discourse. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, pages 19–27, Boulder, Colorado.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. Wit^3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation*, pages 261–268, Trento, Italy.

Zhengxian Gong, Min Zhang, and Guodong Zhou. 2011. Cache-based document-level Statistical Machine Translation. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 909–919, Edinburgh, UK.

Eva Hasler, Barry Haddow, and Philipp Koehn. 2014. Dynamic topic adaptation for smt using distributional profiles. In *Proceedings of the 9th Workshop on Statistical Machine Translation*, pages 445–456, Baltimore, MD, USA.

Laura Mascarell, Mark Fishel, Natalia Korchagina, and Martin Volk. 2014. Enforcing consistent translation of german compound coreferences. In *Proceedings of the 12th Konvens Conference*, pages 58–65, Hildesheim, Germany.

Fandong Meng, Deyi Xiong, Wenbin Jiang, and Qun Liu. 2014. Modeling term translation for document-informed machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 546–556, Doha, Qatar.

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

Xiao Pu, Laura Mascarell, Andrei Popescu-Belis, Mark Fishel, Ngoc-Quang Luong, and Martin Volk. 2015. Leveraging compounds to improve noun phrase translation from chinese and german. In *Proceedings of the ACL-IJCNLP 2015 Student Research Workshop*, pages 8–15, Beijing, China.

Ronald Rosenfeld. 1994. A hybrid approach to adaptive statistical language modeling. In *Proceedings of the Workshop on Human Language Technology*, pages 76–81, Stroudsburg, PA, USA.

Jörg Tiedemann. 2010. Context adaptation in Statistical Machine Translation using models with exponentially decaying cache. In *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*, pages 8–15, Uppsala, Sweden.

Tong Xiao, Jingbo Zhu, Shujie Yao, and Hao Zhang. 2011. Document-level consistency verification in Machine Translation. In *Proceedings of the 13th Machine Translation Summit*, pages 131–138, Xiamen, China.

Deyi Xiong and Min Zhang. 2013. A topic-based coherence model for statistical machine translation. In *Proceedings of the 27th AAAI Conference on Artificial Intelligence*, Bellevue, Washington, USA.

Deyi Xiong and Min Zhang. 2014. A sense-based translation model for statistical machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 1459–1469, Baltimore, Maryland.