Chapter 13

Learning to translate

Timo Honkela, Mathias Creutz, Tiina Lindh-Knuutila, Sami Virpioja, Jaakko J. Väyrynen

13.1 Introduction

Learning to translate research focuses on developing methods and tools facilitating translations between different languages and even between different dialects or domains. Underpinning development of learning to translate methodology is the fact that contextual, experiential and/or disciplinary diversity impede interpersonal communication and understanding. Our research focuses on the use of unsupervised statistical machine learning. However, in comparison with the traditional approach in statistical machine translation (SMT), we want to take into account known linguistic levels and theories. This does not take place by encoding linguistic knowledge manually to the systems but through architectural choices. For instance, the basic statistical machine translation approach does not properly take into account the morphological or the semantic level. These issues are discussed in the following.

We have applied a method of unsupervised morphology learning to a state-of-the-art phrase-based SMT system [2]. In SMT, words are traditionally used as the smallest units of translation. Such a system generalizes poorly to word forms that do not occur in the training data. In particular, this is problematic for languages that are highly compounding, highly inflecting, or both. An alternative way is to use sub-word units, such as morphemes. We have used the Morfessor algorithm to find statistical morpheme-like units (called morphs) that can be used to reduce the size of the lexicon and improve the ability to generalize. This approach is described more in detail in Section 13.3.

The more general the domain or complex the style of the text the more difficult it is to reach high quality translation. The same applies to natural language understanding. All systems need to deal with problems that relate to the lack of semantic coverage and understanding of the pragmatic level of language. Statistical machine translation systems typically rely on applying Bayes' rule:

> We assign to every pair of strings, s (source) and t (target), in two languages a number $P(t|s)$, which is the probability that a translator, when presented with s, will produce t as the translation. Using Bayes' theorem, one can write $P(s|t) = P(t|s) * P(s)/P(t)$

Thus, in the basic SMT approach, the inputs and outputs are handled only as strings of symbols (consider, e.g., [1, 3]. The system does not receive or deal with information on the meaning of the expressions. To overcome this limitation, there are a number of systems with a hybrid approach, using, for instance, a parser that annotates the training samples with (syntactic and) semantic labels. However, as we wish to minimize the manual effort in development machine translation systems, we have chosen not to use traditional parsers or labeling schemes. Rather, we build on distributional information. Namely, the finding that word co-occurrence statistics, as extracted from text corpora, can provide a natural basis for semantic representations has been gaining growing attention. Words with similar distributional properties often have similar semantic properties. Therefore, it is possible to dynamically build semantic representations of the lexical space through the statistical analysis of the contexts in which words co-occur. In addition to distributional information on the word occurrences in text corpora, also other kinds of contextual information may be used.

The early work by Ritter and Kohonen with artificially generated short sentences as well as contextual information showed the feasibility of the approach outlined above [4]. This work was extended to natural data in [2]. In Section 13.4, we describe how the use of the Self-Organizing Map can be extended to multilingual processing in order to find semantic grounding for expressions in multiple languages. Some preliminary results on *visual grounding of meaning* are also discussed.

Before addressing the use of unsupervised learning in finding morphological and semantic models that are useful for machine translation, we consider in the following the structural complexity of a number of European languages in Section 13.2. The basic motivation for this analysis lies in the hypothesis that the translation between such two languages is relatively easier that encode information in a similar manner with respect to morphology and syntax. The results of the analysis should thus help in designing translation strategies for automated solutions for various pairs of languages.

References

[1] P. F. Brown, S. A. D. Pietra, V. J. D. Pietra, and R. L. Mercer (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, vol. 19(2), pp. 263–311.

[2] T. Honkela, V. Pulkki, and T. Kohonen (1995). Contextual relations of words in Grimm tales, analyzed by self-organizing map. In *Proceedings of ICANN'95, International Conference on Artificial Neural Networks*, vol. II, pp. 3–7. Nanterre, France: EC2.

[3] P. Koehn, F. J. Och, and D. Marcu (2003). Statistical phrase-based translation. In *Proceedings of NAACL'03, North American Chapter of the Association for Computational Linguistics on Human Language Technology.* pp. 48–54. Morristown, NJ, USA: Association for Computational Linguistics.

[4] H. Ritter and T. Kohonen, (1989). Self-organizing semantic maps. *Biological Cybernetics*, vol. 61, no. 4, pp. 241–254.

13.2 Analyzing structural complexity of languages

The European Union has 21 official languages (including Irish from 1st of January 2007), which have approximately 407 million speakers. We have analyzed parallel corpora in these 21 languages using statistical, unsupervised learning methods to study the similarities and differences of the languages in different levels. We have compared these results with traditional linguistic categorizations like division into language groups, morphological complexity and syntactic complexity [3]. The aim of the study has been to evaluate the possibility of using statistical methods in different tasks related to statistical machine translation. For instance, for some language pairs the issues related to morphological analysis may be particularly relevant. For some other language pairs, one may have to pay particular attention to the word order. These kinds of questions can be taken into account when the statistical models to be used are chosen.

Use of compression as a measure for complexity is based on the concept of Kolmogorov complexity. Informally, for any sequence of symbols, the Kolmogorov complexity of the sequence is the length of the shortest algorithm that will exactly generate the sequence and then stop. In other words, the more predictable the sequence, the shorter the algorithm needed is and thus the Kolmogorov complexity of the sequence is also lower [4]. Kolmogorov complexity is uncomputable, but file compression programs can be used to estimate the Kolmogorov complexity of a given file. A decompression program and a compressed file can be used to (re)generate the original string. A more complex string (in the sense of Kolmogorov complexity) will be less compressible. Estimations of complexity using compression has been used for different purposes in many areas. Juola [2] introduces comparison of complexity between languages on morphological level for linguistic purposes.

To get a meaningful interpretation for the order of languages in the word order complexity counting, linguistic literature was consulted for independent figures. Bakker [1] has analyzed flexibility of language's word order, which is based on 10 factors, such as order of verb and object in the language, order of adjective and its head noun, order of genitive and its head noun, etc. The flexibility of the language in Bakker's counting can be given with a numeric value from 0 to 1: if the flexibility figure is close to zero, the language is more inflexible in its word order, if the figure is closer to one, the language is more flexible in its word order. In the information theoretic framework of the compression approach flexibility and inflexibility can be interpreted naturally as higher and lower degrees of complexity, i.e. predictability. In the table below, figures based on Bakker's counting of the flexibility values for the individual languages are given together with values given by compression analysis.

If one compares the figures given by Bakker in column 3 to figures given by compression based calculation in column 6, we can see, that the overall order of the languages based on these independent calculations converge well. The lower end of the scale is quite analogous in both analyses consisting of five same languages with differences in the order. There are also some differences in the orders given by the two analyses. The syntactic complexity of Lithuanian seems to be estimated higher by compression than by Bakker's flexibility value (rank 16 vs. 8). Slovene has also a higher flexibility value than its complexity value (rank 14 vs. 7). Greek is also higher in Bakker's counting than in complexity analysis (rank 17 vs. 11). In our compression calculations Finnish and Estonian are estimated almost equally complex, but in Bakker's analysis Estonian is less complex than Finnish (rank 18 vs. 13).[3]

Bakker's results			Compression results		
1.	fr	0.10	1.	fr	0.66
2.	ga	0.20	2.	es	0.68
3.	es	0.30	3.	pt	0.68
4.	pt	0.30	4.	ga	0.69
5.	it	0.30	5.	it	0.69
6.	da	0.30	6.	en	0.69
7.	mt	0.30	7.	sl	0.71
8.	lt	0.30	8.	nl	0.71
9.	en	0.40	9.	mt	0.72
10.	nl	0.40	10.	da	0.72
11.	de	0.40	11.	el	0.73
12.	sv	0.40	12.	sv	0.75
13.	et	0.40	13.	lv	0.75
14.	sl	0.50	14.	de	0.75
15.	lv	0.50	15.	pl	0.76
16.	sk	0.50	16.	lt	0.76
17.	el	0.60	17.	sk	0.77
18.	pl	0.60	18.	et	0.78
19.	fi	0.60	19.	fi	0.79

References

[1] D. Bakker (1998). Flexibility and Consistency in Word Order Patterns in the Languages of Europe. In Siewierska, A. (ed.): *Constituent Order in the Languages of Europe. Empirical Approaches to Language Typology.* pp. 381–419. Mouton de Gruyter, Berlin New York.

[2] P. Juola (1998) Measuring Linguistic Complexity: the Morphological Tier. *Journal of Quantitative Linguistics*, vol. 5, pp. 206–213.

[3] K. Kettunen, M. Sadeniemi, T. Lindh-Knuutila and T. Honkela (2006). Analysis of EU Languages Through Text Compression. In *Proceedings of FinTAL, the 5th International Conference on NLP*, pp. 99–109. Turku, Finland, August 23–25.

[4] M. Li, X. Chen, X. Li, B. Ma, P. M. B. Vitányi (2004). The Similarity Metric. *IEEE Transactions on Information Theory*, vol. 50, pp. 3250–3264.

13.3 Morphology-Aware Statistical Machine Translation

Statistical machine translation was applied to the direct translation between eleven European languages, all those present in the Europarl corpus, by [1]. An impressive number of 110 different translation systems were created, one for each language pair. Koehn discovered that the most difficult language to translate from or to is Finnish. Finnish is a non-Indo-European language and is well known for its extremely rich morphology. As verbs and nouns can, in theory, have hundreds and even thousands of word forms, data sparsity and out-of-vocabulary words present a huge problem even when large corpora are available.

It appears that especially translating into a morphologically rich language poses an even more substantial problem than translating from such a language. The study also showed that English, which has almost exclusively been used as the target language, was the easiest language to translate into. Thus it is natural to suspect that English as a target language has biased SMT research.

In the following, we describe how we have used morphological information found in an unsupervised manner in SMT [2]. We have tested the approach with the three Nordic languages, i.e., Finnish, Danish and Swedish. Danish and Swedish are closely related languages but differ considerably from Finnish. Danish and Swedish are grammatically very close and much of the vocabulary is shared except for some differences in pronunciation and orthography.

The parallel Europarl corpus [1] of European Parliament Proceedings was used to train our models. Word segmentation models for both source and target languages were trained using Morfessor. At this point, two data set were created for each alignment pair: one with the original word tokens and the other with morph tokens. This allowed us to create a comparable baseline system. We used standard state-of-the-art n-gram language models trained with the target language text. Phrase-based translation models were trained with Moses, an open-source statistical machine translation toolkit [3]. A phrase-based system translates short segments of consecutive words in contrast to word-based translation, which translates one word at a time. Phrases enable more natural language generation and flexibility in translating, for instance, idioms, collocations, inflected words forms and compound words in which the number of words may not stay the same across translation. The parameters of word-based and morph-based system were the same, except for the maximum number of tokens in a phrase, that was higher with morph-based systems to cover approximately the same number of words than word-based systems.

Figure 13.1 shows an example how in addition to the the different tokens, morph-based translation first segments words into morphs and finally after the translation constructs words from the translated morphs. Having morph tokens lowers the type counts greatly compared to words. The segmentation naturally increases tokens counts slightly. Reduced type counts help with sparse data, and this was especially prominent with Finnish. On the other hand, increased token counts seem to make the word alignment and translation process more complicated.

In the word-based translation model, only the words that were present in the training data can be translated. The other words are left untranslated, even though they may simply by an inflected form of a known word. Thus we expected to get less untranslated words with the morph-based system. This was true, as shown in Table. 13.1. An examination of the untranslated words reveals that a higher number of compound words and inflected word forms are left untranslated by the word-based systems.

As in most of the recent studies, we have used the BLEU scores [4] for quantitative evaluation. BLEU is based on the co-occurrence of n-grams between a produced transla-

a	flera reglerande åtgärder behöver införas .			
b	flera	reglerande åtgärder	behöver införas	.
c	eräitä	sääntelytoimia	on toteutettava	.
d	eräitä sääntelytoimia on toteutettava .			

e	flera reglerande åtgärder behöver införas .				
f	flera_0 reglera^*_0 nde_+ åtgärd^*_0 er_+ behöv^*_0 er_+ in^*_- föra^*_0 s_+ $._0$				
g	flera_0	reglera^*_0 nde_+	åtgärd^*_0 er_+	behöv^*_0 er_+ in^*_- föra^*_0 s_+	$._0$
h	erä^*_0 itä_+	sääntely^*_0	toimi^*_0 a_+	on_0 toteute^*_0 tta^*_+ va_+	$._0$
i	erä^*_0 itä_+ sääntely^*_0 toimi^*_0 a_+ on_0 toteute^*_0 tta^*_+ va_+ $._0$				
j	eräitä sääntelytoimia on toteutettava .				

Figure 13.1: Examples of word-based and morph-based Finnish translations for the Swedish sentence "Flera reglerande åtgärder behöver införas ." (*Several regulations need to be implemented .*) The top figure shows the word-based translation process with the source sentence (a), the phrases used (b) and their corresponding translations (c), as well as the final hypothesis (d). The bottom figure illustrates the morph-based translation process with the source sentence as words (e) and as morphs (f), the morph phrases used (g) and their corresponding translations (h), as well as the final hypothesis with morphs (i) and words (j). Each morph if either a prefix (−), a stem (0) or a suffix (+), marked by the lowerscript. A superscript (∗) marks the morphs that are not the last one in the word.

word / morph	→ Danish	→ Finnish	→ Swedish
Danish →		128 / 31	74 / 12
Finnish →	189 / 41		195 / 44
Swedish →	76 / 21	132 / 42	

Table 13.1: Number of sentences with untranslated words out of 1 000 with word-based and morph-based phrases.

tion and a reference translation. BLEU score has been critized, for instance, as in some cases human evaluation gives grossly different results. It is also clear that for morphologically rich languages, such as Finnish, it is harder to get good scores on a word-token based evaluation method. In Table 13.2, the differences between the scores for word-based and morph-based systems are shown, with statistically significant differences highlighted. According to these results, the translations based on morph phrases were slightly worse, but only in two cases the decrease was statistically significant.

References

[1] P. Koehn (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT Summit X, 10th Machine Translation Summit*, pp. 79–86. Phuket, Thailand, Sep 13–15.

[2] S. Virpioja, J. J. Väyrynen, M. Creutz and M. Sadeniemi (2007). Morphology-Aware Statistical Machine Translation Based on Morphs Induced in an Unsupervised Manner. In *Proceedings of MT Summit XI, 11th Machine Translation Summit*, pp. 491–498. Copenhagen, Denmark, Sep 10–14.

	→ Danish	→ Finnish	→ Swedish
Danish →		-0.60	-0.52
Finnish →	-1.23		**-2.14**
Swedish →	-0.46	**-1.14**	

Table 13.2: Absolute changes in BLEU scores from word-based translations to morph-based translations. The maximum phrase length was 7 for words and 10 for morphs. 4-gram language models were used for both. Statistically significant differences are marked with boldface fonts.

[3] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, & E. Herbst (2007). Moses: Open source toolkit for statistical machine translation. In *Annual Meeting of ACL, demonstration session*. Czech Republic, June.

[4] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu (2002). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of ACL 2002*, pp. 311–318. Morristown, NJ, USA.

13.4 Self-Organizing Semantic Representations for Machine Translation

Discussing the fundamental problems of translation, Quine has presented a situation in which one is confronted with a situation in which one must attempt to make sense of the utterances and gestures that the members of a previously unknown tribe make [3]. Quine claimed that it is impossible, in such a situation, to be absolutely certain of the meaning that a speaker of the tribe's language attaches to an utterance. For example, if a speaker sees a rabbit and says "gavagai", is she referring to the whole rabbit, to a specific part of the rabbit, or to a temporal aspect related to the rabbit. Even further, if one considers the symbol grounding problem [1], there can practically even be an infinite number of conceptualizations of the situation. Maybe the members of the tribe not only consider the whole rabbit or some parts or aspects of it as potentially relevant points of reference but, e.g., due to their cultural context they consider some other patterns of perception. Namely, considering the complex pattern recognition process, it is far from trivial to create a perception of a rabbit from the raw visual and auditory input.

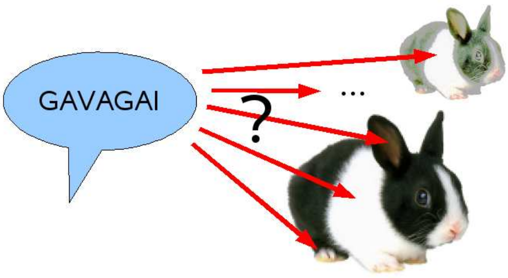

Figure 13.2: An illustration of the reference problem (see text for details).

Quine mentions that one can form manuals of translation [3]. The observer examines the utterances as parts of the overall linguistic behavior of the individual, and then uses these observations to interpret the meaning of all other utterances. Quine continues that there will be many such manuals of translation since the reference relationship is indeterminate. He allows that simplicity considerations not only can be used to choose between competing manuals of translation but that there is even a remote possibility of getting rid of all but one manual.

It seems that propositional logic as the underlying epistemological framework unnecessarily complicates the consideration. For Quine it was necessary to consider a number of logically distinct manual of translation hypotheses. However, if one considers the issue within the framework of statistics, probability theory and continuous multidimensional representations of knowledge, one can consider the conditional probability of different hypotheses and partial solutions which do not need to be logically coherent. Moreover, the search for translation mappings can be seen as a process that may (or may not) converge over time. For Quine meaning is not something that is associated with a single word or sentence, but is rather something that can only be attributed to a whole language. The resulting view is called semantic holism. In a similar fashion, the self-organizing map specifies a holistic conceptual space. The meaning of a word is not based on some definition but is the emergent result of a number of encounters in which a word is perceived or used in some context. Moreover, the emergent prototypes on the map are not isolated instances but they influence each other in the adaptive formation process.

Finding a mapping between vocabularies of two different languages, the results of a new experiment are reported in the following. Maps of words are often constructed using distributional information of the words as input data. The result is that the more similar the contexts in which two words appear in the text, the closer the words tend to be on the map. We have extended this basic idea to cover the notion of context in general. We have considered the use of a collection of words in two languages, English and German, in a number of contexts. In this experiment, the contexts were real-life situations rather than some textual contexts.

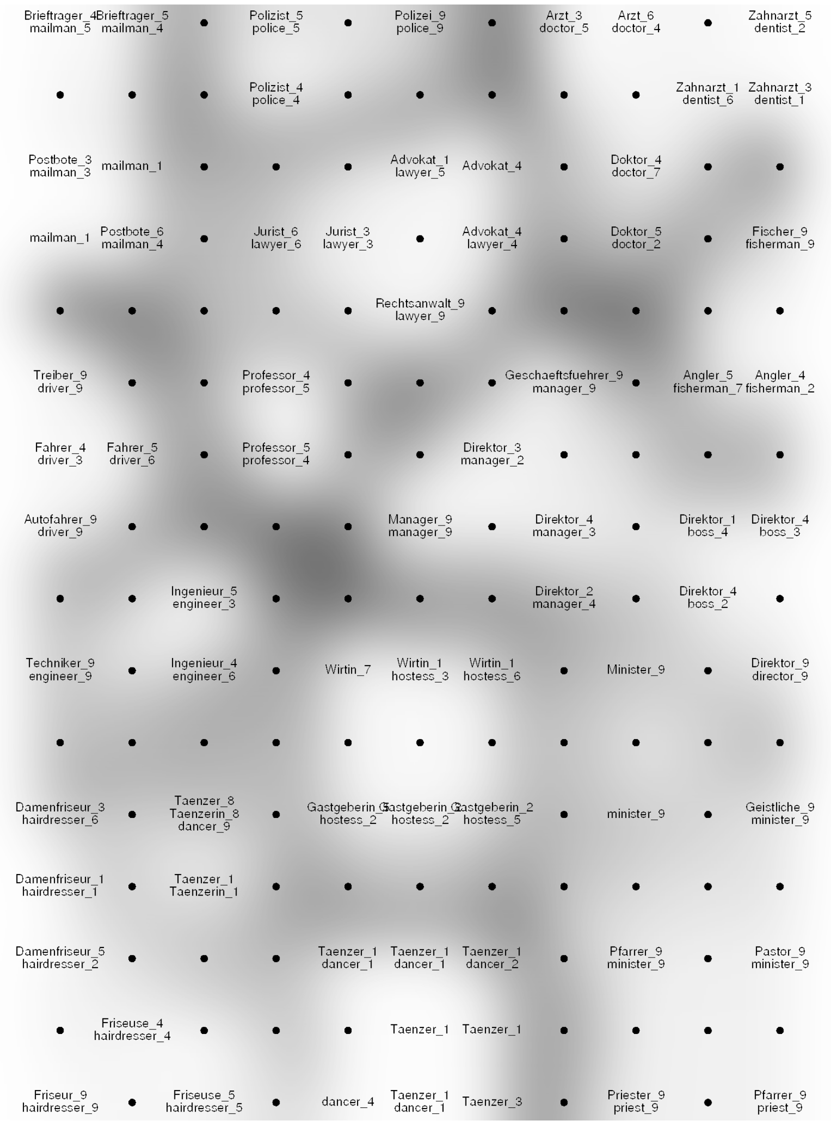

Figure 13.3: An illustration of the reference problem (see text for details).

Figure 13.3 presents the order of some words on a self-organizing map that serves simultaneously two purposes. First, it has organized different contexts to create a semantic landscape. Second, the map includes a mapping between the English and German words used in the analysis. The input for the map consists of words and their contexts. The German vocabulary includes 33 words (Advokat, Angler, Arzt, Autofahrer, ..., Zahnarzt) and the English vocabulary of 17 words (boss, dancer, dentist, director, ..., professor). For each word, there is a assessment by 10 to 27 subjects indicating the degree of suitability for the word to be used in a particular context. The number of contexts used was 19.

The map shows that those words in the two languages that have similar meaning are close to each on the map. In this particular experiment, the German subjects were usually using a larger vocabulary. Therefore, in many areas of the map, a particular conceptual area is covered by one English word (for instance, "doctor" or "hairdresser") and by two or more German words (for instance, "Arzt" and "Doktor" or "Friseur", "Friseuse" and "Damenfriseur").

The research on content-based information retrieval and analysis (see Chapter 7) provides a solid basis for future research in "translation through images". Some initial experiments show that names of concrete objects in different languages can be mapped with each other without any intermediate linguistic/symbolic representation (see [5] for details). In general, these results support the idea of symbol grounding [4].

References

[1] S. Harnad (1990). The symbol grounding problem. *Physica D*, vol. 42, pp. 335–346.

[2] T. Honkela (2007). Philosophical Aspects of Neural, Probabilistic and Fuzzy Modeling of Language Use and Translation In *Proceedings of IJCNN 2007, International Joint Conference on Neural Networks.* Orlando, Florida, Aug 12–17.

[3] W. Quine, (1960). *Word and Object.* MIT Press.

[4] M. Sjöberg, J. Laaksonen, M. Pöllä, T. Honkela (2006). Retrieval of Multimedia Objects by Combining Semantic Information from Visual and Textual Descriptors. In *Proceedings of ICANN 2006, International Conference on Artificial Neural Networks*, pp. 75–83. Athens, Greece.

[5] M. Sjöberg, V. Viitaniemi, J. Laaksonen, T. Honkela (2006). Analysis of Semantic Information Available in an Image Collection Augmented with Auxiliary Data. In *Proceedings of IFIP, Conference on Artificial Intelligence Applications and Innovations*, pp. 600–608. Athens, Greece.