Adaptive Translation: Finding Interlingual Mappings using Self-Organizing Maps

Timo Honkela, Sami Virpioja and Jaakko Väyrynen

Adaptive Informatics Research Centre
Helsinki University of Technology
P.O. Box 5400, FI-02015 TKK
Espoo, Finland

Abstract. This paper presents a method for creating interlingual word-to-word or phrase-to-phrase mappings between any two languages using the self-organizing map algorithm. The method can be used as a component in a statistical machine translation system. The conceptual space created by the self-organizing map serves as a kind of interlingual representation. The specific problems of machine translation are discussed in some detail. The proposed method serves in alleviating two problems. The main problem addressed here is the fact that different languages divide the conceptual space differently. The approach can also help in dealing with lexical ambiguity.

1 Introduction

In the following, the area of machine translation is discussed in broad terms. Specific problems that make high quality machine translation a difficult task are described.

1.1 Natural Language Understanding and Machine Translation

The language of a person is idiosyncratic and based on the subjective experiences of the individual. For instance, two persons may have a different conceptual or terminological density of the topic under consideration. A layperson, for instance, is likely to describe a phenomenon in general terms whereas an expert uses more specific terms. Moore and Carling [1] state that languages are in some respect like maps. If each of us sees the world from our particular perspective, then an individual's language is, in a sense, like a map of their world. Trying to understand another person is like trying to read the map of the other, a map of the world from another perspective [1]. If some persons speak the same language, many symbols in their vocabularies are the same. However, as discussed above,

one cannot assume that the vocabularies of any two agents are exactly the same. These issues related to natural language understanding are even more challenging if communication in multiple languages is considered.

The term *machine translation* (MT) refers to computerized systems responsible for the production of translations with or without human assistance. It excludes computer-based translation tools which support translators by providing access to on-line dictionaries, remote terminology databanks, transmission and reception of texts, etc.

Development of high quality MT has proved to be a very challenging task. At best, the translation results can be reasonably good, for example, when the domain is small or a lot resources is used in developing the system. At worst, the results can be practically incomprehensible.

Another important challenging factor is that the development of a machine translation system requires a *lot of human effort*. This is true especially in the case of systems that are based on human-specified rules for morphology, syntax, lexical selection, semantic analysis and generation.

1.2 Translation Challenges

In the following, a list of seven *more detailed problems* are given. These problems need to be addressed in the development of MT systems and in comparisons of MT systems and approaches. A more philosophical consideration on the problems of translation is provided in [2].

1. Differences in conceptual mapping: Different languages divide the conceptual space differently, for example in the domain of physical space (cf. e.g. [3]) or colors (cf. e.g. [4]). The possibility of exact meaning preserving translation from one language to another can even be questioned on these grounds. In addition to overlapping areas of reference related to two words in different languages, there may be also words missing in some language for some particular phenomenon. One can, of course, introduce new words into a language in order to facilitate a one-to-one mapping between concepts in two languages would be an impossible task. Therefore, from the semantic and pragmatic point of view, it seems that both human and machine translation are always subject to some inaccuracy and there can be arguably several good ways to translate one expression to another language. This fact makes the evaluation of machine translation systems more difficult. For statistical MT this might also mean that it would be useful to have several different parallel translations available for a source language expression in order to facilitate modeling of the variety.

2. Lexical and syntactic ambiguity: In an MT system, one has to determine a suitable interpretation or to find the correct analysis of the syntactical structure for an expression in order to be able to find the corresponding expression in another language. This process is usually denoted as disambiguation. Traditionally lexical disambiguation refers to a process in which the "correct" meaning among several alternatives is determined. However, one can also consider the possibility that the underlying space of interpretations can be continuous (cf.

e.g. [5]). In that case the disambiguation process would determine a more specific distribution or point in the conceptual space. However, it is nowadays much more commonplace in the MT systems to consider a word to have one or several distinct meanings.

3. Word, multi-word and sentence alignment: Meaning is often conveyed by phrases rather than by single words. In one language, an idea may be expressed with one separate word, in another with a compound word, and in a third one with a multi-word phrase (collocation). It is also possible that there are one-to-many or many-to-many relationships in the sentence level. In an extreme case, one word in the source language might require several sentences in the target language in order to express approximately the same meaning in the cultural context of each language. It may also be that in one language some particular word classes are in use while in another language they are not. For instance, articles are used in many languages, in some others not. The alignment problem is a very actively researched topic (cf. e.g. [6]).

4. Differences in word order: Even for laypersons, a well known difference between languages is related to the word order. It is commonplace to categorize a language to have one basic word order, referring to the fact in which order subject, verb and object appear in a declarative sentence. In many languages, changes in word order occur due to topicalization or in questions. The basic word order is unmarked, i.e., it contains no extra information to the listener. In addition to the differences in the order of the basic constituents in a simple declarative sentence, there are many other points of difference, e.g., the order in which a noun and its attributes appear.

5. Inflectional word forms: Some languages have complex morphology, i.e., they have have inflectional word forms. The inflections may denote, e.g., the case of a noun or the tense of a verb. The existence of complex word forms relates closely to the alignment problem. In translation, each segment (morphemes) of one word form in the source language may correspond to one or several words in the target language. The order of the segments may very well be different from the order of the words (consider also the previous point).

6. Text segmentation: In some languages the words are not explicitly separated from each other in a text. This introduces another challenge: Before or associated with the alignment process, the system has to detect the word borders.

7. Speech-to-text and text-to-speech transformation: In the case of MT between spoken languages, the system has to perform transformation between the modalities unless the mapping is conducted directly without any textual intermediate representation. We are not aware of any such approach even if we consider it in principle possible.

Corpus-based methods have been introduced in order to deal, for instance, with the knowledge acquisition bottleneck. The availability of large corpora makes it possible to use various probabilistic and statistical methods to acquire automatically lexicon entries, parsing and disambiguation rules and other useful

representations for a translation system. Next we describe the basic statistical translation model that is currently widely used.

1.3 Bayesian Models in Translation

Bayes' rule tells that

$$p(A|B) = p(B|A)p(A)/p(B), \tag{1}$$

where $p(A)$ is the prior probability or marginal probability of A, and $p(A|B)$ is the conditional probability of A, given B. Let us consider a situation in which we wish to translate Finnish sentences, f, into English sentences, e. Bayes' rule gives $p(e|f)p(f) = p(e, f) = p(f|e)p(e)$ and reduces to the basic equation of statistical machine translation: maximize $p(e|f) = p(f|e)p(e)$ over the appropriate e. This splits the translation problem into a translation model ($p(f|e)$) and a language model ($p(e)$). The decoding algorithm, given these models and a new sentence f, finds translation e.

In their classical paper, Peter Brown and his colleagues described a series of five statistical models of the translation process and gave algorithms for estimating the parameters of these models applying the basic equation [7]. During recent years, this statistical approach has had considerable successes, based on the availability of large parallel corpora and some further methodological developments (consider, e.g., [8, 9]).

Translation or mapping between two language through use of the self-organizing map can be a viable solution for the problems 1 and 2 described above. The traditional Bayesian approach does not take these issues into account. Construction of maps of words based on the self-organizing map algorithm is next presented as an introductory theme. The self-organizing map is a natural means to build a conceptual space that can be used as a link between two or several languages.

1.4 Self-Organizing Map

The Self-Organizing Map (SOM) [10, 11] defines an ordered mapping, a kind of projection from a set of given data items onto a regular, usually two-dimensional grid. A model m_i is associated with each grid node. These models are computed by the SOM algorithm. A data item will be mapped into the node whose model is most similar to the data item, i.e., has the smallest distance from the data item in some metric. The model is then usually a certain weighted local average of the given data items in the data space. But in addition to that, when the models are computed by the SOM algorithm, they are more similar at the nearby nodes than between nodes located farther away from each other on the grid. In this way the set of the models can be regarded to constitute a similarity graph, and structured 'skeleton' of the distribution of the given data items. [12]

1.5 Maps of Words

Charniak [13] presents the following scheme for grouping or clustering words into classes that reflect the commonality of some property.

- Define the properties that are taken into account and can be given a numerical value.
- Create a vector of length n with n numerical values for each item to be classified.
- Cluster the points that are near each other in the n-dimensional space.

Handling computerized form of written language rests on processing of discrete symbols. How can a symbolic input such as a word be given to a numeric algorithm? One useful numerical representation can be obtained by taking into account the sentential context in which the words occur. Before utilization of the context information, however, the numerical value of the code should not imply any order to the words. Therefore, it will be necessary to use uncorrelated vectors for encoding. The simplest method to introduce uncorrelated codes is to assign a unit vector for each word. When all different word forms in the input material are listed, a code vector can be defined to have as many components as there are word forms in the list. This method, however, is only practical in small experiments. With a vocabulary picked from a even reasonably large corpus the dimensionality of the vectors would become intolerably high. If the vocabulary is large, the word forms can be encoded by quasi-orthogonal random vectors of a much smaller dimensionality [14]. Such random vectors can still be considered to be sufficiently dissimilar mutually and not to convey any information about the meaning of the words. Mathematical analysis of the random encoding of the word vectors is presented in [15].

2 Experiments with Interlingual Mapping

To illustrate the idea of using the Self-Organizing Map in finding a mapping between vocabularies of two different languages, the results of two new experiments are reported in the following.

2.1 Situation context

Like discussed in earlier sections, the maps of words are often constructed using the sentential contexts of 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. Here this basic idea is extended to cover the notion of context in general: We consider 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 1 presents the order of a number of words on a self-organizing map that serves simultaneously two purposes. First, it has organized different contexts to

create a conceptual landscape (see, e.g., [5]). 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 32 words (Advokat, Angler, Arzt, Autofahrer, ..., Zahnarzt) and the English vocabulary of 16 words (boss, dancer, dentist, director, etc.). 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 is 19. The resulting map is shown in Figure 1.

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"). It is important to notice that the model covers both translation between languages and within languages. Namely, rather than dealing with German and English, the same model can be built for the language used in medical contexts by experts and laypersons.

2.2 Textual context

In this experiment, a map of words was constructed using real-life sentential contexts of words as input data. The vocabulary was bilingual, as were the sentential contexts. The idea is to get words with similar contexts to appear close to each other on the map. With a bilingual vocabulary, we wish to find 1) a semantic ordering of words in the map, and 2) possible translation pairs of words from those close to each other on the map.

The words and their sentential contexts were obtained from the sentence-aligned English–German part of the parallel Europarl corpus (version 3) [16]. Each context spanned the words in the aligned regions in the corpus, covering 1,298,966 sentence pairs.

The input for the map consists of English and German words and their bilingual contexts. The most frequent 150 nouns were selected separately from both languages to get a vocabulary of 300 words. Note that the vocabulary does not contain sensible translations for all of its words. The contexts for the selected 300 words were calculated from the most frequent 3889 words in the two languages. In contrast to the previous experiment with feature contexts, each word was represented only with a single context vector.

The context variables were calculated as the number of co-occurrence counts of the words in the vocabulary with the context words, resulting with 300 word vectors with 3889 context variables. The variances of the context variables were normalized to one to make all the variables equally important. The L2 norm of the word vectors was normalized to one, which makes the Euclidian metric used by the SOM very closely related to the cosine similarity typically used with vector space models. The resulting map is shown in Figure 2 with the 300 words positioned to the closest map units.

The map shows the bilingual vocabulary evenly distributed to the map, with translations of words next to each other. The words are also ordered semantically, with related words close to each other (e.g., 'world', 'europe', 'country'). Inflections of the same word are typically also close to each other (e.g., 'Jahr', 'Jahre', 'Jahren'), with sometimes singular and plural forms a bit more separate. The map also shows several possible translations next to each other (e.g., 'aid', 'help', 'Hilfe'). Words without translation equivalents in the vocabulary are located to semantically near words (e.g., 'Lösung' near 'problem').

The results reported above are in an interesting contrast with another study in which words of two languages were presented in linguistic contexts [17]. Li and Farkas found out that the two languages were strictly separated on the map. The differences are explained by the selection of the word contexts: Li and Farkas used the distributions of preceeding and following words in bilingual, intermixed sentences as semantic context, whereas we use "bag-of-words" representation of the full sentences in both languages.

3 Conclusions and Discussion

The main difference between the approach outlined in the previous sections and the Bayesian method, in its commonly used form, is that the semantic or conceptual space is explicitly modeled in the SOM-based approach. Thus, the mapping between any two languages is based on an intermediate level of representation. This approach resembles, to some degree, the idea of using a knowledge-based interlingua in machine translation. The underlying philosophical assumptions about knowledge are, however, quite different. In a knowledge-based interlingua, the semantics of natural language expressions are typically represented as propositions and relations in symbolic hierarchical structures. The SOM can be used to span a continuous and multidimensional conceptual space in a data-driven manner. Moreover, the approach provides a natural means to deal with multimodal data [18] and, thus, deal with the symbol grounding problem [19].

References

1. Moore, T., Carling, C.: The Limitations of Language. Macmillan Press, Houndmills (1988)
2. Honkela, T.: Philosophical aspects of neural, probabilistic and fuzzy modeling of language use and translation. In: Proceedings of IJCNN'07, International Joint Conference on Neural Networks. (2007)
3. Bowerman, M.: The origins of children's spatial semantic categories: cognitive versus linguistic determinants. In Gumperz, J., Levinson, S.C., eds.: Rethinking linguistic relativity. Cambridge University Press, Cambridge (1996) 145–76
4. Berlin, B., Kay, P.: Basic Color Terms: Their Universality and Evolution. University of California Press (1991 (1969))
5. Gärdenfors, P.: Conceptual Spaces. MIT Press (2000)
6. Ahrenberg, L., Andersson, M., Merkel, M.: A simple hybrid aligner for generating lexical correspondences in parallel texts. In: Proceedings of COLING-ACL'98. (1992) 29–35

7. Brown, P.F., Pietra, S.A.D., Pietra, V.J.D., Mercer, R.L.: The mathematics of statistical machine translation: Parameter estimation. Computational Linguistics **19(2)** (1993) 263–311
8. Koehn, P., Och, F.J., Marcu, D.: Statistical phrase-based translation. In: NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, Morristown, NJ, USA, Association for Computational Linguistics (2003) 48–54
9. Zhang, R., Yamamoto, H., Paul, M., Okuma, H., Yasuda, K., Lepage, Y., Denoual, E., Mochihashi, D., Finch, A., Sumita, E.: The NiCT-ATR Statistical Machine Translation System for IWSLT 2006. In: Proceedings of the International Workshop on Spoken Language Translation, Kyoto, Japan (2006) 83–90
10. Kohonen, T.: Self-organizing formation of topologically correct feature maps. Biological Cybernetics **43**(1) (1982) 59–69
11. Kohonen, T.: Self-Organizing Maps. Volume 30 of Springer Series in Information Sciences. Springer, Berlin, Heidelberg (2001)
12. Kohonen, T., Honkela, T.: Kohonen network. Scholarpedia (2007) 7421
13. Charniak, E.: Statistical Language Learning. MIT Press, Cambridge, Massachusetts (1993)
14. Ritter, H., Kohonen, T.: Self-organizing semantic maps. Biological Cybernetics **61**(4) (1989) 241–254
15. Kaski, S.: Dimensionality reduction by random mapping: Fast similarity computation for clustering. In: Proceedings of IJCNN'98, International Joint Conference on Neural Networks. Volume 1. IEEE Service Center, Piscataway, NJ (1998) 413–418
16. Koehn, P.: Europarl: A parallel corpus for statistical machine translation. In: Proceedings of the 10th Machine Translation Summit, Phuket, Thailand (2005) 79–86
17. Li, P., Farkas, I.: A self-organizing connectionist model of bilingual processing. In: Bilingual sentence processing, North-Holland (2002) 59–85
18. Laaksonen, J., Viitaniemi, V.: Emergence of ontological relations from visual data with self-organizing maps. In: Proceedings of SCAI 2006, Scandinavian Conference on Artificial Intelligence, Espoo, Finland (2006) 31–38
19. Harnad, S.: The symbol grounding problem. Physica D **42** (1990) 335–346

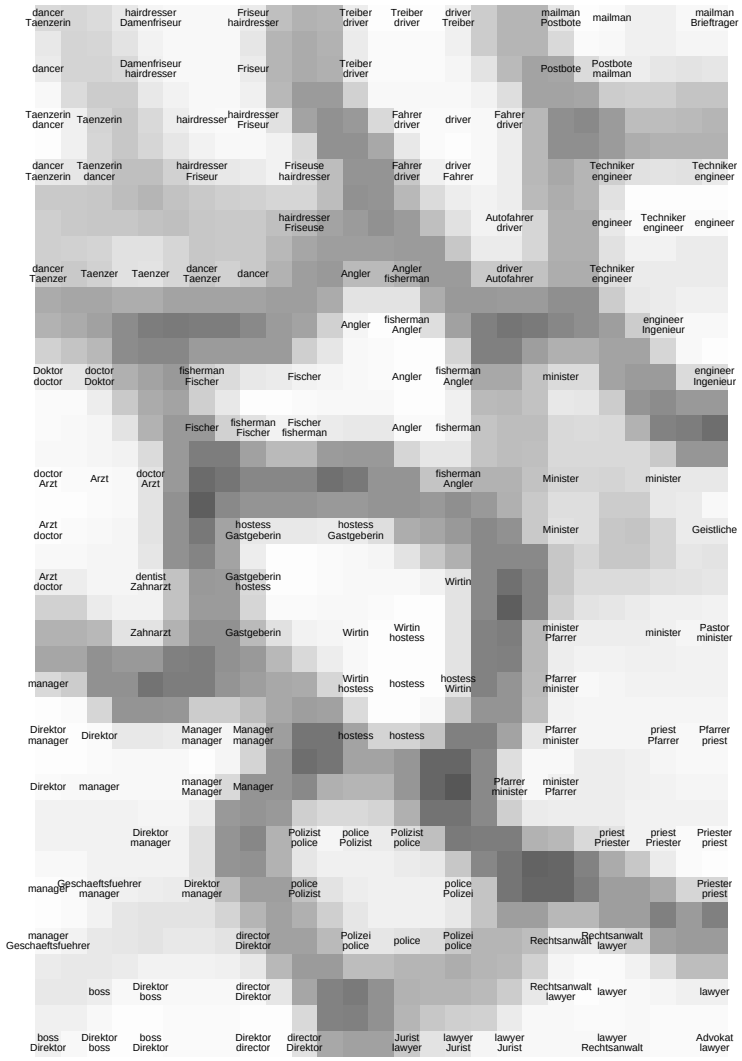

Fig. 1. A collection of German and English words positioned a conceptual landscape based on a Self-Organizing Map of contexts. The darker the shade of gray, the longer are the distances in the original input space. Thus, relatively light areas correspond to conceptual areas or clusters. The dots on the map denote empty prototypes, i.e., model vectors that are not the best match of any of the words under consideration.

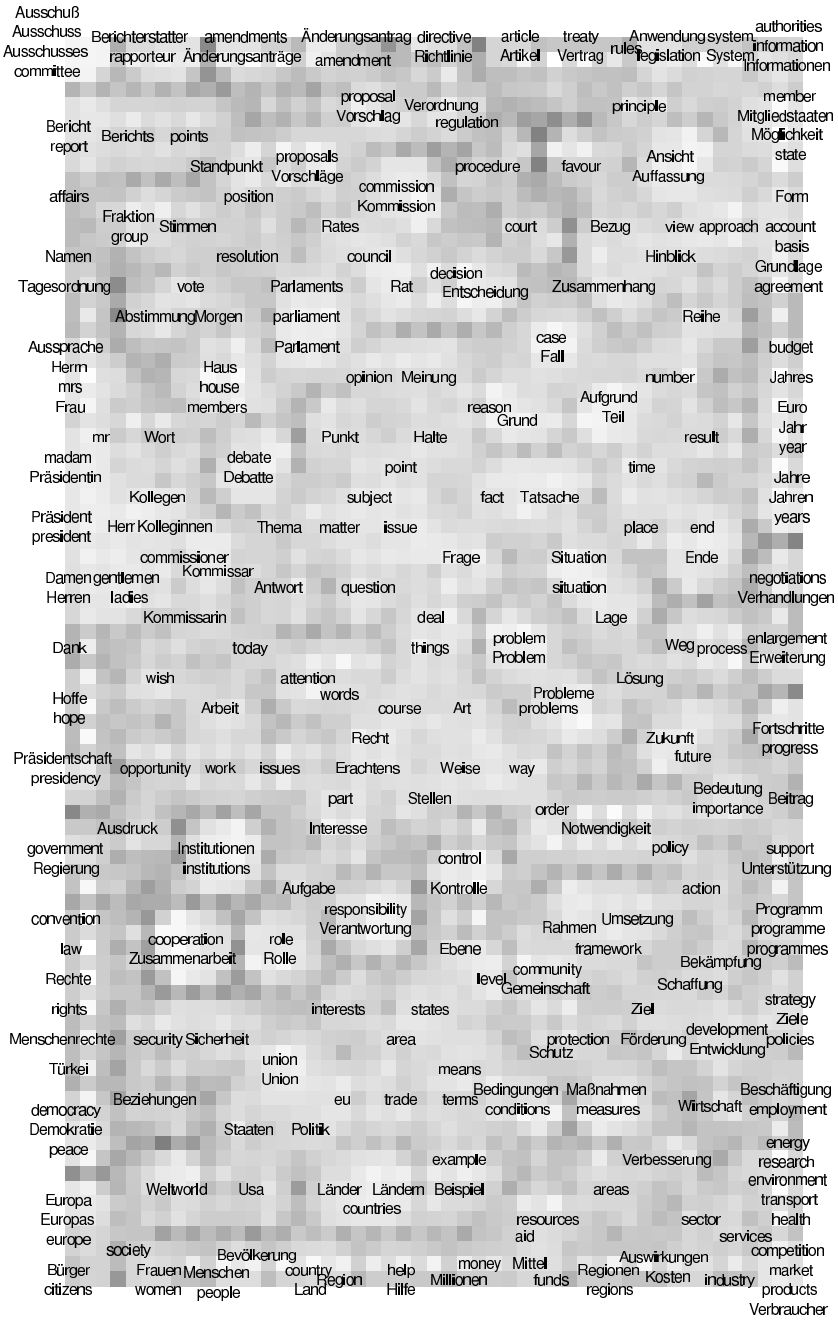

Fig. 2. German and English words on a Self-Organizing Map of contexts. The interpretation of the map is similar to that of Figure 1, except the map was taught with real-life bilingual sentential contexts, and for each word there is only one vector.