Translational equivalence in Statistical Machine Translation or meaning as co-occurrence

Lieve Macken and Els Lefever
University College Ghent - Ghent University

In this paper, we will describe the current state-of-the-art of Statistical Machine Translation (SMT), and reflect on how SMT handles meaning. Statistical Machine Translation is a corpus-based approach to MT: it derives the required knowledge to generate new translations from corpora. General-purpose SMT systems do not use any formal semantic representation. Instead, they directly extract translationally equivalent words or word sequences – expressions with the same meaning – from bilingual parallel corpora. All statistical translation models are based on the idea of word alignment, i.e., the automatic linking of corresponding words in parallel texts. The first generation SMT systems were word-based. From a linguistic point of view, the major problem with word-based systems is that the meaning of a word is often ambiguous, and is determined by its context. Current state-of-the-art SMT-systems try to capture the local contextual dependencies by using phrases instead of words as units of translation. In order to solve more complex ambiguity problems (where a broader text scope or even domain information is needed), a Word Sense Disambiguation (WSD) module is integrated in the Machine Translation environment.

1. Introduction: Statistical Machine Translation

Statistical Machine Translation (SMT) is one of the best performing corpus-based approaches to natural language processing (NLP). Unlike the work that has been carried out in the field of language philosophy, corpus-based approaches traditionally have not tried to define the meaning of a word in a philosophical or explicit way, but restricted themselves to the study of meaning in context. Other machine translation approaches, such as the Interlingua approach, have tried to formally represent the meaning of words (see section 2).

The idea of linking the meaning of a word to its context has a long history that starts with the distributional theory of meaning, which links the meaning of a word to its distribution and further states that two words are distributionally similar if they appear in similar contexts. This theory of meaning goes back to Harris' Distributional Hypothesis (Harris 1968), suggesting a direct link between distributional similarity and semantic similarity: two words that tend to occur in similar contexts tend to have similar meanings.

This idea is also exploited by lexicographers today, who use corpus evidence for creating dictionaries. For each dictionary entry, a KWIC (keyword in context search) is performed, and all relevant meanings and patterns of use are distilled from the example sentences.

As SMT only performs a shallow analysis of the context, it is confronted with a number of inherent problems of the nature of language. First, there are semantic and structural differences between languages. Examples of these differences are non-compositional expressions (e.g. Dutch *in het oog springend (prominent)*), words that can only be translated by multiword paraphrases (e.g. the Portuguese word *saudade*), structural differences (e.g. the English verb in *I like swimming* is translated in German by means of an adverb: *Ich schwimme gern*), etc. The second category of problems is caused by ambiguity, which is still considered to be the most fundamental problem of language technology. For SMT, we refer to ambiguity whenever there is uncertainty about the meaning of a word or sentence in a text. The origin of the ambiguity can be morphological (e.g. German compound *Staubecken* can be translated as *water reservoir* (stau-becken) or as *dust corners* (staub-Ecken)), syntactic (e.g. *John saw the boy with the telescope*, where it is not clear whether it is John who used the telescope or the boy), semantic (homographs and polysemes (e.g. *Turn the truck to the right* versus *the deposits turn into sludge, Bush became president following the 2000 presidential election* versus *the raging bush and forest fires*)) or referential (*Mary hit her bag against the vase, it broke,* where you need real world knowledge to know that *it* refers to the *vase* as a bag does not break).

In language technology, ambiguity is partially resolved by performing automatic analysis on all linguistic levels (tokenisation, part-of-speech tagging, parsing, semantic analysis, anaphora resolution, etc). The remainder of the paper shows how different generations of Machine Translation systems have tackled the major problems MT is confronted with.

2. Machine Translation

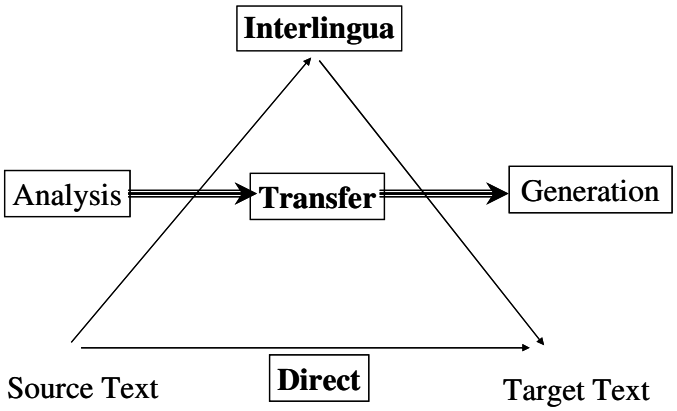

Figure 1: Vauquois triangle

The ultimate goal of Machine Translation, whichever approach is taken, is to translate a source language text into a target language text in a fully automatic way. The different approaches can be classified according to the level of linguistic analysis that is used, which is often illustrated by means of the Vauquois triangle (Vauquois 1968).

In *direct* Machine Translation systems, the translation of each word or phrase is looked up in a dictionary and the source language word is substituted by the equivalent word or phrase in the target language. Dictionary lookups may be done on the word form or on the lemmatized form.[1] So, in the direct approach only a minimum of linguistic analysis is used. In *transfer*-based approaches, the source language text is syntactically analysed (parsed), and the source language parse structure is transformed into a target language parse structure. The target language sentence is generated from the target language parse structure. In *interlingua* approaches, the source language text is analysed into some abstract meaning representation,[2] called an interlingua. The target language sentence is generated from this abstract meaning representation.

It goes without saying that 'perfect' translation can only be achieved by using semantic knowledge, which involves a deep syntactic and semantic analysis of the text. Machine Translation is still considered as one of the most difficult problems in the field of Natural Language Processing, and the problem of automatically producing a high quality translation of an arbitrary text is far too hard to automate completely (Jurafsky & Martin 2000: 800). Nevertheless, most state-of-the-art systems are based on a superficial analysis, and can produce reasonable quality.

While the earlier approaches to Machine Translation (and Natural Language Processing in general) were *rule-based*, the current approaches are *data-driven*, which means that all the knowledge needed for translation is extracted from parallel texts. The statistical approaches to Machine Translation are also data-driven. In most cases, parallel corpora are aligned at sentence level, meaning that sentences of the source texts are connected with sentences of the target texts.

3. Statistical Machine Translation

Two inherent characteristics of good translations form the basis of the architecture of SMT systems:[3]

- Translations preserve the meaning of the source language (*faithfulness* or *fidelity*);
- translations are as natural as an utterance of the target language (*naturalness* or *fluency*).

The goal of Statistical Machine Translations is to produce an output that maximizes these two factors. In order to achieve this goal, an SMT system must be able (1) to quantify faithfulness, (2) to quantify fluency and (3) needs an algorithm that finds the sentence that maximizes the product of these two factors (Jurafsky & Martin 2000: 819).

Fluency is measured by probabilistic monolingual language models, which are in most cases n-gram models. The probability of an n-gram (i.e. in case of a 3-gram, the probability that a sequence of three words occurs) is derived from large monolingual corpora of the target language.

Since this paper deals with meaning, the first problem, "how to quantify faithfulness or fidelity" is more of interest to us. The central question we will try to answer is: **How do existing SMT systems measure how close the meaning of a source sentence is to the meaning of a translated sentence?** The basic factor often used in metrics of fidelity is the degree to which all words in the target sentence are plausible translations of the words in the source sentence. Thus, the probability of a sentence being a good translation can be approximated as the product of the probabilities that each target language word is an appropriate translation of some source language word (Jurafsky & Martin 2000: 821).

In order to calculate the probability of a sentence and its translation, the system therefore needs to know for every target language word, the probability of its mapping to every source language word.
These translation probabilities are derived from parallel texts (aligned source-target sentence pairs). The problem of deriving the translation probabilities from parallel texts is closely related to the problem of word alignment, which is explained in the next section.

4. Word Alignment

Statistical word alignment is an unsupervised method, which means that it starts from unannotated (raw) data from a large sentence-aligned corpus. It is based on the assumption of co-occurrence: words that are translations of each other co-occur more often than random in aligned sentence pairs. The output of a statistical word alignment model is a large bilingual word list with probability estimations.

The most widely used statistical word alignment models are the IBM translation models (Brown et al. 1993). The simplest IBM model – IBM Translation Model One – is a purely lexical model: it only takes into account word frequencies in source and target sentences. The higher numbered IBM Models build on IBM Model One and take into account word order (distortion) and model the probability that a source word aligns to n target words (fertility). The IBM models allow only 1:n word mappings. A detailed description and comparison of the IBM models can be found in Och and Ney (2003).

It was already mentioned in section 2 that SMT does not make use of hard-coded rules, but uses probabilistic knowledge sources in the form of probability distributions. The statistical word alignment process is also guided by probabilities. Suppose one has a corpus that is manually aligned at the word level. In order to extract the alignment probabilities one could just count how many times a source word is translated by a certain target word. Unfortunately, large corpora in which word alignments are manually indicated do not exist and are time-consuming to create. Therefore, a methodology was developed to estimate these probabilities without human intervention.

One technique to estimate the translation probabilities is the Expectation Maximization (EM) algorithm (Manning & Schütze 1999: 488). The EM algorithm is a learning method that iteratively carries out two steps: in the first step (Expectation step), probabilities are assigned to all word pairs by applying a word alignment model; in the second step (Maximization step), the model is adapted based on new counts collected for all word pairs.

To start up the process, the initial word alignment model applies a uniform distribution: it assumes that all correspondences between source and target words in an aligned sentence pair are equally likely. In the subsequent iterations, it calculates the relative frequencies, which it uses as a model in a subsequent iteration.

The EM process is illustrated in Figure 2. In the Expectation step of the first iteration, all source words of each sentence are aligned to all target words of that sentence. After a few iterations, the system knows (based on the frequency counts) that the connections between *the* and *de*, and *doctor* and *dokter* are more likely. This is indicated in Figure 2 by means of a thicker line. In most systems, four or five iterations are used.

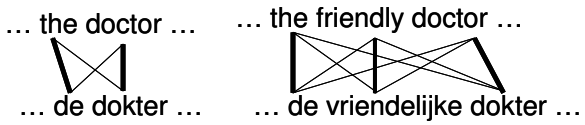

Figure 2: illustration of the EM process

To illustrate how SMT arrives at meaning we used the Perl implementation of IBM Model one that is part of the Microsoft Bilingual Sentence Aligner (Moore 2002), on a 10-million-word sentence-aligned English-Dutch sample of the Europarl corpus (Koehn 2005).

Table 1: estimated probability distributions for all Dutch translations of *doctor* and for all Dutch translations of *head* based on Europarl

p(arts|doctor)	0.481	p(hoofd|head)	0.269
p(doktor|doctor)	0.082	p(kop|head)	0.068
p(medische|doctor)	0.017	p(regeringsleider|head)	0.019
p(ingrepen|doctor)	0.015	p(voorsprong|head)	0.014
p(huisarts|doctor)	0.015	p(hoofdelijke|head)	0.014
p(ingrijpt|doctor)	0.012	p(inwoner|head)	0.014

From the resulting bilingual dictionary, we selected all translations of *doctor* and all translations of *head* with a probability value higher than 0.01. The resulting word pairs and probabilities are given in Table 1. A first observation is that the two most probable translations in the Europarl corpus are *arts* (48%) and *doktor* (8%). A second observation is that other semantically related words that often co-occur in the target sentence (e.g. *medische*, *ingrepen*, *huisarts*) also get a higher probability value even if they have no equivalent meaning. However, the difference between the probability values of these pairs and those of the two most probable translations is considerable: the lowest value of the latter is nearly five times higher than the highest value of the former, which can be considered as less reliable.

Similar observations can be made for the translations of *head*. The two most probable translations are *hoofd* (27%) and *kop* (7%). The Dutch translation *regeringsleider* for *head* reveals a problem of the IBM models: they are asymmetric. They can only model 1:n correspondences as they take as starting point the source word and estimate conditional probabilities (i.e.

the probability that a target word is a translation of a source word, given the source word). Multiword units (the Dutch word *regeringsleider* corresponds with *Head of Government*) are problematic for the word-based models, as every word (*Head*, *of* and *Government*) is treated as a separate entry. To overcome this problem, the IBM models are often used in two directions: from source to target and from target to source.

We explained in section 3, that the problem of quantifying faithfulness or fidelity for a sentence and its translation is approximated as the product of the probabilities that each target language word is an appropriate translation of some source language word. This is in fact an oversimplification of the problem, as a word-for-word translation is assumed.

In most translations, however, translational correspondences are more complex, and only for some words can word-by-word correspondences be found. The rest of the sentence is translated on the level of combinations of words. The following example shows a sentence pair where more complex translational correspondences (*challenge – daag uit*, *in the secrecy of – achter de gesloten deuren van*, *Council Chamber – Raadskamer*) are indicated manually. Such more complex correspondences (see also example 1, below) remain problematic for the word-based SMT systems.

(1) En: I challenge any minister who may resist these proposals in the secrecy of the Council chamber to ...
Nl: Ik daag iedere minister die deze voorstellen achter de gesloten deuren van de Raadskamer verwerpt uit om ...

Table 2: manually indicated translational correspondences

I	*Ik*
challenge	*daag...uit*
any	*iedere*
minister	*minister*
who	*die*
may resist	*verwerpt*
these	*deze*
proposals	*voorstellen*
in the secrecy of	*achter de gesloten deuren van*
the	*de*
Council chamber	*Raadskamer*
to	*om*
...	*...*

In spite of the fact that statistical word alignment systems are able to extract automatically translational equivalences at word level and estimate

probability distributions for the translation pairs, which can be used to indicate the "relative importance" of the different translations, they face a number of limitations.

- Most word alignment systems start from plain text corpora, which means that they align word forms (word tokens). A useful abstraction can be achieved by lemmatizing the corpus prior to word alignment. Lemmatization can be a workaround to solve the data scarcity problem if the parallel corpus is small.
- A related problem is the treatment of semantically related words as unrelated tokens (e.g. *act* and *action*). This problem is more difficult to solve.
- Generally speaking, words are defined as space-delimited tokens. However, complex words are formed differently in different languages: For example, English and Dutch use a different compounding strategy (e.g. *regeringsleider – Head of Government*, *anti-terrorism policy – antiterrorismebeleid*). The word alignment system has problems to cope with such differences.
- During word alignment no contextual information is used.[4] Often, the correct translation depends on the context. In the first-generation word-based statistical machine translation systems, context information was only available in the n-gram language models, which only code monolingual information of the target language. Therefore, the second-generation SMT systems work with larger units that are translationally equivalent, viz. phrases.

5. Phrase-based statistical Machine Translation

A first attempt to improve Word-based SMT-systems by adding contextual information to the translation models is the use of a phrase translation table. In a phrase translation table translations of word pairs or phrases are stored in such a way that the immediate local context can be used to determine the translational equivalence. On the basis of word alignments, current phrase-based SMT systems automatically extract bilingual phrases. As we have seen in section 4, the IBM models are asymmetric. To overcome this problem, the IBM models are run in two directions: from source to target and from target to source. Different symmetrisation heuristics can be used to combine the word alignments of both translation directions (Och & Ney 2003).

Figure 3 shows the output of Moses, an open source phrase-based SMT system (Koehn et al. 2007), after symmetrisation of the alignment points. Please observe that the system is not error-free: the English words *may* and *resist* have not been aligned, and *to* has been erroneously aligned with the Dutch word *om*.

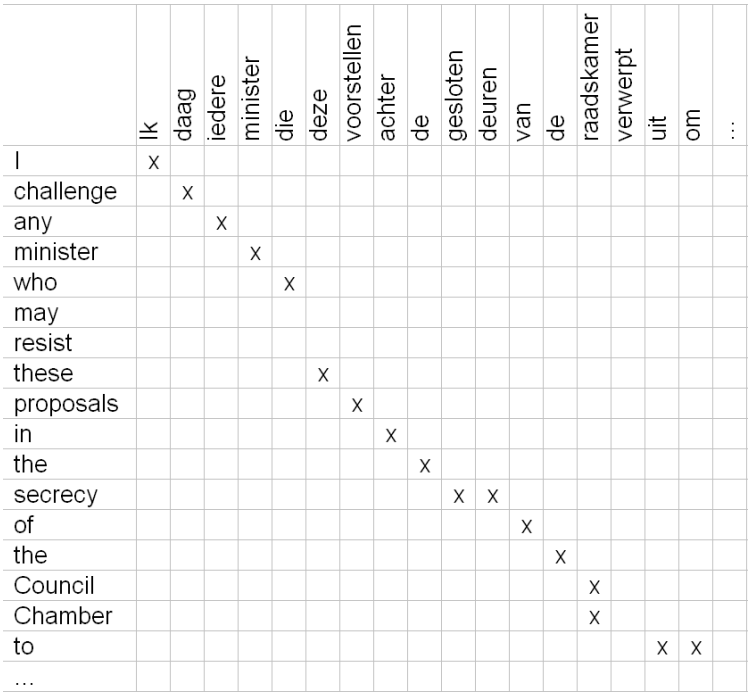

Figure 3: symmetrised word alignment points indicated by x's

On the basis of the symmetrised alignment points, the bilingual phrases are extracted automatically and stored in the so-called phrase translation table. Any phrase pair that is consistent with the symmetrised word alignment is collected, where *consistent* is defined as: the words in the phrase pair are aligned to each other and not to any words outside the phrase pair (Koehn et al. 2005).

In the example used for Figure 3, the following phrases have been extracted: (I challenge/ Ik daag), (challenge any/ daag iedere), (any minister/ iedere minister), (minister who/ minister die), (these proposals/ deze voorstellen), (proposals in/ voorstellen achter), (in the/ achter de), (the secrecy/ de gesloten deuren), (secrecy of/ gesloten deuren van), (of the/ van de), (the Council Chamber/ de raadskamer), (to/ uit om), (I challenge any/ Ik daag iedere), (challenge any minister/ daag iedere minister), etc.

The phrase translation tables as defined above allow the phrase-based SMT systems to capture certain translational phenomena as long as they are contiguous chunks. The limitation that only contiguous chunks are included in the phrase table is especially problematic for languages such as Dutch and German. Since they contain a high percentage of separable prefix verbs and adopt a less strict word order (in comparison to English), the prefix is often separated over a long distance from the verb (e.g. *daag ... uit* in the example sentence above). Macken (2007) demonstrated that espe-

cially for Dutch non-contiguous correspondences account for 2.5 to 5% of all alignments. As far as we know, only Simard et al. (2005) allow non-contiguous phrases in an SMT system.

Although the phrase-based SMT systems perform significantly better than the word-based systems, they still face a lot of problems. First, they cannot capture the movement of hierarchical structures during translation. Several attempts have been made to include syntactic knowledge into statistical MT systems. However, as their main objective is to model word order problems, we will not pursue this issue in detail in this paper. Second, phrase-based MT systems only handle ambiguity in case it can be resolved by incorporating the immediate context. To solve the more complex ambiguity problems, where a broader text scope or even domain information is needed, a real disambiguation module must be integrated into the machine translation environment. This novel approach to MT is described in the next section.

6. Word Sense Disambiguation in Statistical Machine Translation

Until recently, semantic ambiguity has only been handled in an implicit way in SMT. In the IBM models, contextual information is very limited: the language model (which quantifies naturalness) uses n-grams, whereas the translation model makes use of a phrase translation table that only captures the immediate local context.

In recent years, small improvements obtained by adding dedicated Word Sense Disambiguation (WSD) modules to the SMT system have been reported. Cabezas and Resnik (2005) have tried to cast the problem of lexical selection in SMT as a WSD problem in which the "senses" are target translations of the source word. Example 2 below shows how their approach improves the general translation quality by producing a correct translation for the ambiguous word "carta" in Spanish:

(2) Source sentence:
señor presidente, he **votado a favor de esta carta** en buena parte por la influencia que nuestro colega ingo friedrich y el profesor herzog han ejercido en su contenido.

Baseline MT output:[5]
*i **voted for this in a letter** to the influence mr ingo friedrich and professor herzog have exercised their content.*

WSD MT output:
*mr president, **voted in favour of the charter** in large part by the influence mr ingo friedrich and professor herzog have exercised their content.*

Carpuat and Wu (2007), however, claim to have demonstrated that lexical semantics are useful for SMT. They have reported significant improvements on the standard metrics that are used for MT evaluation by adding a disambiguation module that is integrated in a phrase-based SMT system.

The following two sections describe how a Word Sense Disambiguation module works.

7. Word Sense Disambiguation

Starting from the hypothesis that the meaning of a word is determined by its context, a number of machine learning approaches have been developed in the NLP field for addressing both the synonymy and polysemy problem that often causes poor machine translation quality. Machine learning algorithms allow computers to 'learn' from training data; in this way inductive methods derive rules and patterns from large amounts of data.

In a similar way, the shallow approaches to Word Sense Disambiguation (WSD or choosing the right sense of a polysemous word in a given context) do not try to understand the text in order to solve the problem either, but use contextual information from the training data.

Agirre and Edmonds (2007) consider WSD as a classification task: a machine learning algorithm assigns the correct class ("sense" in this case) to a word, based on its context. As each polysemous word gets assigned its own classes or senses, it will require the construction of its own classifier. This is not the case for other classification tasks, such as Part-of-Speech tagging, where the class inventory is fixed (and consists of a predefined set of Part-of-Speech tags) and where only one classifier is trained an applied on random input text.

For WSD, the classes that can be predicted correspond to the different senses of the polysemous word. If we consider the following two sentences with the ambiguous word "bank", in example 3, the first one will get the class/sense label "SHORE", wherease the second will get the class/sense label "FINANCE":

(3) Today the **banks** of four rivers are polluted (Label: SHORE)
The Australian **bank** issues blunt warnings on interest rates (Label: FINANCE)

As mentioned before, the number of classes per ambiguous word corresponds to the predefined sense inventory (list of all senses) per word. In order to construct sense inventories for polysemous words, most WSD algorithms make use of digital lexical resources such as WordNet. These resources, however, have two fundamental problems: the sense distinctions are too detailed for real applications (problem of granularity) and they are

mainly available for English, and to a limited extent only for a couple of other languages (e.g. EuroWordNet). Very fine sense distinctions do not only make the classification task much harder, but also cause an explosion of combinatorial effects (e.g. the example given by Slator and Wilks (1987): "there is a huge envelope of air around the surface of the earth" produced 284,592 sense combinations based on the Longman Dictionary of Contemporary English).

Although WSD researchers have long been convinced that the level of sense-discrimination (distinction of different meanings of a word) needed by NLP corresponds to homographs, they now concede that for some applications one should make finer distinctions. Consequently, the sense inventory is strongly related to the application WSD is used for (e.g. *mouse* is always translated as *souris* in French, thus the difference in meaning might be irrelevant for MT, but is certainly important for other applications such as Information Retrieval where you want to retrieve the documents for "mouse=animal" or "mouse=computer accessory", but not the entire set of documents for both meanings of the word). Recent research in WSD has tried to overcome these difficulties by exploiting parallel corpora for the construction of the sense inventories, the underlying idea being that different senses are often lexicalized differently in other languages. Ideally, languages from different language families are taken into account in order to maximize the cross-linguistic lexicalization of the different senses.

8. Word Sense Disambiguation approaches

WSD classifiers can either be supervised (trained on a corpus of words tagged with their word senses, usually retrieved from a sense inventory such as WordNet) or unsupervised (senses of a given word are distinguished by grouping similar contexts of the ambiguous word). Semi-supervised methods make use of annotated corpora in a bootstrapping process where training examples are used for training a classifier that tags 'sure' (the classifier output having a very high confidence score) unseen occurrences, which are then added to the training corpus.

The first WSD algorithms used to rely on knowledge-based resources (dictionaries, thesauri and lexical knowledge bases). A well-known algorithm is the (Simplified) Lesk Algorithm, which identifies senses of words in context measuring the overlap between the sense definitions of the word and the current context. The following example (Lesk 1986) shows how the algorithm works. To determine the correct sense of *cone* and *pine* in the collocation *pine cone*, the algorithm first looks up the sense definitions of both words.

(4) Pine
(1) seven kinds of **evergreen tree** with needle-shaped leaves
(2) pine
(3) waste away through sorrow or illness
(4) pine for something, pine to do something

Cone
(1) solid body which narrows to a point
(2) something of this shape, whether solid or hollow
(3) fruit of certain **evergreen trees** (fir, pine)

The Lesk algorithm will select sense (1) for pine and sense (3) for cone as these definitions have most words in common. However, the knowledge-based algorithms only achieve a moderate performance (between 50 and 70 percent) which can both be explained by the fact that sense distinctions are very fine (which is a problem) and by (remediable) shortcomings of the resources: dictionary definitions are often too short and use other words to describe the same concept. To overcome some of these problems, more recent algorithms prefer large untagged or annotated corpora as training material. These corpora offer a vast amount of traning examples and a variety of contexts for ambiguous words. Moreover, the sense labels are usually coarse-grained and therefore better suited for training learning algorithms.

Up to now supervised algorithms have proved to be the most successful for tackling WSD. Because of the lack of manually sense-tagged data required for performing supervised learning, here again the exploitation of parallel corpora is gaining ground. Ng et al. (2003) have done experiments for Chinese-English to automatically acquire training data from parallel corpora. Given the word-aligned parallel corpus, the different translations of the ambiguous words serve as "sense tags" for the ambiguous words in the source language, and afterwards the different classes (translations) are mapped to existing WordNet senses. In this way the training examples are enriched with automatically acquired examples.

The idea of using the different translations instead of explicit sense labels has been further developed for elaborating unsupervised approaches. The free availability of Europarl, a corpus of parallel text in 11 languages containing the proceedings of the European Parliament, has also speeded up the exploitation of parallel corpora. The unsupervised approaches do not only use the parallel corpora to provide training examples, but use the translations themselves as "sense classes" for each ambiguous word. Another advantage of using parallel text for constructing the sense inventory is that the corpus can be made domain or application specific. The unsupervised approach seems to work well for specific applications such as Machine Translation or Information Retrieval. Unfortunately, so far no sense inventory for general-purpose WSD has been created.

State-of-the art WSD classifiers typically use a wide range of contextual knowledge to decide on the right label (sense) of an ambiguous word. Context is defined in a broad sense going from collocations (surrounding words) and co-occurring words (bag of words extracted from a window of the X preceding and following sentences) to text genre or topic/domain of the text. Other useful information can be morpho-syntactical (part-of-speech tags, lemma, syntactic dependency relations, etc.) or semantic (semantic class, roles, etc). Research results (Agirre & Edmonds 2007: 233-234) have shown that nouns typically ask for wide context and local collocation information, whereas verbs benefit most from syntactic features.

In order to train the classifier, all relevant information (coded as "features") is extracted for each instance from an ambiguous word and is stored in one feature vector per instance. To classify new occurrences of ambiguous words, the same feature vector is constructed and compared to all training instances in order to find the best corresponding training example and accompanying label.

Table 3 illustrates the extraction of the (simplified) feature vectors for our two *bank* examples in the training corpus:

- Today/Adv the/Det **banks/SHORE** of/Prep four/Num rivers/Noun are/Aux polluted/Participle (Ex 1)
- The/Det Australian/Adj **bank/FINANCE** issues/Verb blunt/Adj warnings/Noun on/Prep interest/Noun rates/Noun (Ex 2)

The first four features that are listed contain Part-of-Speech information (PoS tags of the two preceding and two following words), the last three features contain co-occurrence information. For each ambiguous word, a set of words that often co-occur with the ambiguous target word, is automatically defined (based on co-occurrence frequency). These context features indicate whether this word, the one that often co-occurs with the ambiguous target word, is present in the sentence ("Yes") or not ("No").

The last column in Table 3 contains the Sense Label that is assigned to the ambiguous word in this particular sentence.

Table 3: example of two feature vectors for the word *bank*

	PoS -2	PoS -1	PoS +1	PoS +2	river	interest	Rate	Sense Label
Ex 1	Adv	Det	Prep	Num	Yes	No	No	"shore"
Ex 2	Det	Adj	Verb	Adj	No	Yes	Yes	"finance"

The way to best integrate WSD in SMT seems to be "phrase sense disambiguation". This approach redefines the classical WSD task to move beyond the single word targets, and generalises to multi-word phrase targets

that match better the phrasal lexical selection requirements of the state-of-the art phrase-based SMT systems. Sense candidates for the ambiguous phrases are then defined by the SMT translation lexicon itself.

9. Conclusion

Statistical Machine Translation (SMT) is a corpus-based approach to MT: it derives the required knowledge to generate new translations from corpora. General-purpose SMT systems do not use any formal semantic representation. Instead, they directly extract translationally equivalent words or word sequences – expressions with the same meaning – from bilingual parallel corpora.

All statistical translation models are based on the idea of word alignment, i.e., the automatic linking of corresponding words in parallel texts. Melamed (1998) decomposed the problem of MT lexicon construction process into two parts:

- What are the possible translations for each source word?
- In what context are the various translations used?

Provided that suitable sentence-aligned parallel corpora are available, word alignment methods can answer the first question. To answer the second question, it is necessary – but this is by no means easy – to move away from the word level (word-to-word translation) and go to the phrase level, so that immediate local context can be used to determine translational equivalence. Another promising and complementary approach to select the appropriate translation is to incorporate WSD-techniques, which use contextual clues taken from the whole sentence (and beyond) to discriminate between different word senses.

Bibliography

Agirre, Eneko, & Philip Edmonds (2007). *Word Sense Disambiguation. Algorithms and Applications*. Place: Springer.

Brown, Peter F., Vincent J. Della Pietra, Stephen A. Della Pietra & Robert L. Mercer (1993). "The Mathematics of Statistical Machine Translation: Parameter Estimation". *Computational Linguistics, 19*(2), 263-311.

Cabezas, Clara & Philip Resnik (2005). *Using wsd techniques for lexical selection in statistical machine translation*. Maryland: Institute for Advanced Computer Studies,.

Carpuat, Marine & Dekai Wu (2007). *Context-Dependent Phrasal Translation Lexicons for Statistical Machine Translation.* Proceedings of the Machine Translation Summit XI, European Association of machine Translation, Copenhagen, Denmark.

Harris, Zelig Sabbetai (1968). *Mathematical structures of language*. New York: Wiley.

Jurafsky, Daniel & James H. Martin (2000). *Speech and language processing*. London: Prentice-Hall International.

Katz, J.J. & J.A. Fodor (1963). "The structure of a semantic theory". *Language 39*, 170-210.

Koehn, Philipp (2004). *Pharaoh, a beam search decoder for phrase-based statistical machine translation models: User manual and description for version 1.2,* August 2004.

Koehn, Philipp (2005). *Europarl: a parallel corpus for statistical machine translation.* Proceedings of the Tenth Machine Translation Summit, Phuket, Thailand.

Koehn, Philipp, Amittai Axelrod, Alexandra Birch, Chris Callison-Burch, Miles Osborne & David Talbot. (2005). *Edinburgh system description for the 2005 IWSLT speech translation evaluation.* Proceedings of the International Workshop on Spoken Language Translation: Evaluation Campaign on Spoken Language Translation (IWSLT 2005), Pittsburgh, PA, USA.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris-Callison-Burch, Mercello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin & Evan Herbst (2007). *Moses: Open Source Toolkit for Statistical Machine Translation.* Proceedings of the ACL 2007 Demo and Poster Sessions, Prague, Czech Republic.

Lesk, Michael (1986). *Automated sense disambiguation using machine-readable dictionaries: How to tell a pine cone from an ice cream cone.* Proceedings of the 1986 ACM SIGDOC Conference, Toronto, Canada.

Macken, Lieve (2007). *Analysis of translational correspondence in view of sub-sentential alignment.* Proceedings of the METIS-II Workshop on New Approaches to Machine Translation, Leuven, Belgium.

Manning, Christopher D. & Hinrich Schütze (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA: Massachusetts Institute of Technology.

Melamed, Dan I. (1998). *Empirical methods for MT lexicon development.* Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas (AMTA '98), Langhorne PA, USA.

Moore, Robert C. (2002). *Fast and accurate sentence alignment of bilingual corpora.* Proceedings of the 5th Conference of the Association for Machine Translation in the Americas, Tiburon, California.

Ng, Hwee Tou, Bin Wang & Yee Seng Chan (2003). *Exploiting parallel texts for word sense disambiguation: An empirical study.* Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL), Sapporo, Japan.

Och, Franz Josef, and Hermann Ney (2003). "A systematic comparison of various statistical alignment models". *Computational Linguistics 29*(1), 19-51.

Simard, Michel, Nicola Cancedda, Bruno Cavestro, Marc Dymetman, Eric Gaussier & Cyril Goutte (2005). *Translating with non-contiguous phrases.* Proceedings of the Human Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT-EMNLP-2005), Vancouver.

Slator, Brian M. & Yorick A. Wilks (1987). *Towards semantic structures from dictionary entries.* Proceedings of the 2nd Annual Rocky Mountain Conference on Artificial Intelligence, Boulder, Colorado.

Toury, Gideon (1995). *Descriptive Translation Studies and Beyond.* Amsterdam/Philadelphia: John Benjamins.

Vauquois, Bernard (1968). "Structures profondes et traduction automatique. Le système du CETA." *Revue Roumaine de Linguistique 13*(2), 105-130.

---

[1] During lemmatization, for each orthographic word, the base form (or canonical form) is generated.

[2] One of the earliest attempts to formalize an abstract meaning representation as a string of features can be found in Katz and Fodor (1963).

[3] In Translation Studies, Toury also discerns this dichotomy in his initial norm and refers to it with the terms *adequacy* and *acceptability*: 'Whereas adherence to source norms determines a translation's adequacy as compared to the source text, subscription to norms originating in the target culture determines its acceptability" (Toury, 1995, pp. 56-57).

[4] The higher numbered IBM models try to model word order, however.

[5] As a baseline, they use the Pharaoh decoder (Koehn, 2004).