A Model of Competence for Corpus-Based Machine Translation

Michael Carl
Institut für Angewandte Informationsforschung,
Martin-Luther-Straße 14,
66111 Saarbrücken, Germany,
carl@iai.uni-sb.de

Abstract

In this paper I elaborate a model of competence for corpus-based machine translation (CBMT) along the lines of the representations used in the translation system. Representations in CBMT-systems can be rich or austere, molecular or holistic and they can be fine-grained or coarse-grained. The paper shows that different CBMT architectures are required dependent on whether a better translation quality or a broader coverage is preferred according to Boitet (1999)'s formula: "Coverage * Quality = K".

1 Introduction

In the machine translation (MT) literature, it has often been argued that translations of natural language texts are valid if and only if the source language text and the target language text have the same meaning cf. e.g. (Nagao, 1989). If we assume that MT systems produce meaningful translations to a certain extent, we must assume that such systems have a notion of the source text meaning to a similar extent. Hence, the translation algorithm together with the data it uses encode a formal model of meaning. Despite 50 years of intense research, there is no existing system that could map arbitrary input texts onto meaning-equivalent output texts. How is that possible?

According to (Dummett, 1975) a theory of meaning is a theory of understanding: having a theory of meaning means that one has a theory of understanding. In linguistic research, texts are described on a number of levels and dimensions each contributing to its understanding and hence to its meaning. Traditionally, the main focus has been on semantic aspects. In this research it is assumed that knowing the propositional structure of a text means to understand it. Under the same premise, research in MT has focused on semantic aspects assuming that texts have the same meaning if they are semantically equivalent.

Recent research in corpus-based MT has different premisses. Corpus-Based Machine Translation (CBMT) systems make use of a set of reference translations on which the translation of a new text is based. In CBMT-systems, it is assumed that the reference translations given to the system in a training phase have equivalence meanings. According to their intelligence, these systems try to figure out of what the meaning invariance consists in the reference text and learn an appropriate source language/target language mapping mechanism. A translation can only be generated if an appropriate example translation is available in the reference text.

An interesting question in CBMT systems is thus: what theory of meaning should the learning process implement in order to generate an appropriate understanding of the source text such that it can be mapped into a meaning equivalent target text? Dummett (Dummett, 1975) suggests a distinction of theories of meaning along the following lines:

- In a *rich* theory of meaning, the knowledge of the concepts is achieved by knowing the features of these concepts. An *austere* theory merely relies upon simple recognition of the shape of the concepts. A rich theory can justify the use of a concept by means of the characteristic features of that concept, whereas an austere theory can justify the use of a concept merely by enumerating all occurrences of the use of that concept.
- A *molecular* theory of meaning derives the understanding of an expression from a finite number of axioms. A *holistic* theory, in contrast, derives the understanding of an expression through its distinction from all other expressions in that language. A molecular theory, therefore, provides criteria to associate a certain meaning to a sentence and can explain the concepts used in the language. In a holistic theory nothing is specified about the knowledge of the language other than in global constraints related to the language as a whole.

In addition, the granularity of concepts seems crucial for CBMT implementations.

- A *fine-grained* theory of meaning derives concepts from single morphemes or separable words of the language, whereas in a *coarse-grained*

theory of meaning, concepts are obtained from morpheme clusters. In a fine-grained theory of meaning, complex concepts can be created by hierarchical composition of their components, whereas in a coarse-grained theory of meaning, complex meanings can only be achieved through a concatenation of concept sequences.

The next three sections discuss the dichotomies of theories of meaning, *rich vs. austere*, *molecular vs. holistic* and *coarse-grained vs. fine-grained* where a few CBMT systems are classified according to the terminology introduced. This leads to a model of competence for CBMT. It appears that translation systems can either be designed to have a broad *coverage* or a high *quality*.

2 Rich vs. Austere CBMT

A common characteristic of all CBMT systems is that the understanding of the translation task is derived from the understanding of the reference translations. The inferred translation knowledge is used in the translation phase to generate new translations.

Collins (1998) distinguishes between Memory-Based MT, i.e. memory heavy, linguistic light and Example-Based MT i.e. memory light and linguistic heavy. While the former systems implement an austere theory of meaning, the latter make use of rich representations.

The most superficial theory of understanding is implemented in purely memory-based MT approaches where learning takes place only by extending the reference text. No abstraction or generalization of the reference examples takes place.

Translation Memories (TMs) are such purely memory based MT-systems. A TM e.g. TRADOS's Translator's Workbench (Heyn, 1996), and STAR's TRANSIT calculates the graphemic similarity of the input text and the source side of the reference translations and return the target string of the most similar translation examples as output. TMs make use of a set of reference translation examples and a (k-nn) retrieval algorithm. They implement an austere theory of meaning because they cannot justify the use of a word other than by looking up all contexts in which the word occurs. They can, however, enumerate all occurrences of a word in the reference text.

The TM distributed by ZERES (Zer, 1997) follows a richer approach. The reference translations and the input sentence to be translated are lemmatized and part-of-speech tagged. The source language sentence is mapped against the reference translations on a surface string level, on a lemma level and on a part-of-speech level. Those example translations which show greatest similarity to the input sentence with respect to the three levels of description are returned as the best available translation.

Example Based Machine Translation (EBMT) systems (Sato and Nagao, 1990; Collins, 1998; Güvenir and Cicekli, 1998; Carl, 1999; Brown, 1997) are richer systems. Translation examples are stored as feature and tree structures. Translation templates are generated which contain - sometimes weighted - connections in those positions where the source language and the target language equivalences are strong. In the translation phase, a multi-layered mapping from the source language into the target language takes place on the level of templates and on the level of fillers.

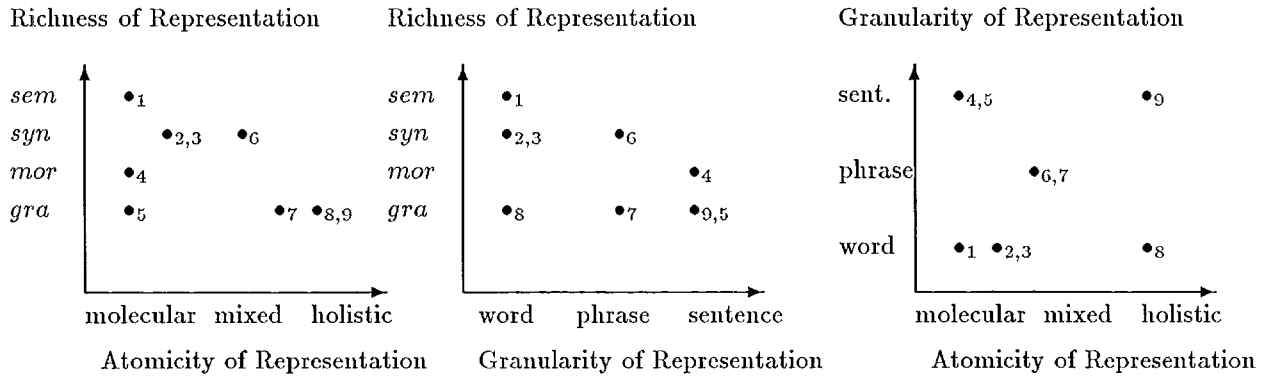
The ReVerb EBMT system (Collins, 1998) performs sub-sentential chunking and seeks to link constituents with the same function in the source and the target language. A source language subject is translated as a target language subject and a source language object as a target language object. In case there is no appropriate translation template available, single words can be replaced as well, at the expense of translation quality.

The EBMT approach described in (Güvenir and Cicekli, 1998) makes use of morphological knowledge and relies on word stems as a basis for translation. Translation templates are generalized from aligned sentences by substituting differences in sentence pairs with variables and leaving the identical substrings unsubstituted. An iterative application of this method generates translation examples and translation templates which serve as the basis for an example based MT system. An understanding consists of extraction of compositionally translatable substrings and the generation of translation templates.

A similar approach is followed in EDGAR (Carl, 1999). Sentences are morphologically analyzed and translation templates are decorated with features. Fillers in translation template slots are constrained to unify with these features. In addition to this, a shallow linguistic formalism is used to percolate features in derivation trees.

Sato and Nagao (1990) proposed still richer representations where syntactically analyzed phrases and sentences are stored in a database. In the translation phase, most similar derivation trees are retrieved from the database and a target language derivation tree is composed from the translated parts. By means of a thesaurus semantically similar lexical items may be exchanged in the derivation trees.

Statistics based MT (SBMT) approaches implement austere theories of meaning. For instance, in Brown et al. (1990) a couple of models are presented starting with simple stochastic translation models getting incrementally more complex and rich by introducing more random variables. No linguistic



•1: Sato and Nagao (1990) •2: EDGAR Carl (1999) •3: GEBMT Güvenir and Cicekli (1998)
•4: ZERES Zer (1997) •5: TRADOS Heyn (1996) •6: ReVerb Collins (1998)
•7: Brown (1997) •8: Brown et al. (1990) •9: McLean (1992)

Figure 1: Atomicity, Granularity and Richness of CBMT

analyses are taken into account in these approaches. However, in further research the authors plan to integrate linguistic knowledge such as inflectional analysis of verbs, nouns and adjectives.

McLean (McLean, 1992) has proposed an austere approach where he uses neural networks (NN) to translate surface strings from English to French. His approach functions similar to TM where the NN is used to classify the sequences of surface word forms according to the examples given in the reference translations. On a small set of examples he shows that NN can successfully be applied for MT.

3 Molecular vs. Holistic CBMT

As discussed in the previous section, all CBMT systems make use of some text dimensions in order to map a source language text into the target language. TMs, for instance, rely on the set of graphemical symbols i.e. the ASCII set. Richer systems use lexical, morphological, syntactic and/or semantic descriptions. The degree to which the set of descriptions is independent from the reference translations determines the molecularity of the theory. The more the descriptions are learned from and thus depend on the reference translations the more the system becomes holistic. Learning descriptions from reference translations makes the system more robust and easy to adjust to a new text domain.

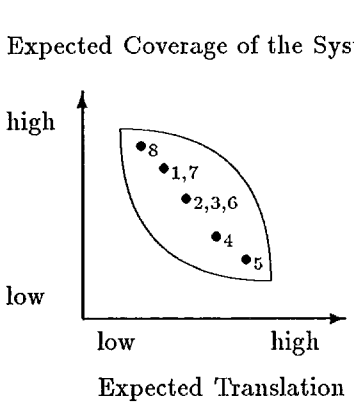
SBMT approaches e.g. (Brown et al., 1990) have a purely holistic view on languages. Every sentence of one language is considered to be a possible translation of any sentence in the other language. No account is given for the equivalence of the source language meaning and the target language meaning other than by means of global considerations concerning frequencies of occurrence in the reference text. In order to compute the most probable translations, each pair of items of the source language and the target language is associated with a certain probability. This prior probability is derived from the reference text. In the translation phase, several target language sequences are considered and the one with the highest posterior probability is then taken to be the translation of the source language string.

Similarly, neural network based CBMT systems (McLean, 1992) are holistic approaches. The training of the weights and the minimization of the classification error relies on the reference text as a whole. Temptations to extract rules from the trained neural networks seek to isolate and make explicit aspects on how the net successfully classifies new sequences. The training process, however, remains holistic.

TMs implement the molecular CBMT approach as they rely on a static distance metric which is independent from the size and content of the case base. TMs are molecular because they rely on a fixed and limited set of graphic symbols. Adding further example translations to the data base does not increase the set of the graphic symbols nor does it modify the distance metric. Learning capacities in TMs are trivial as their only way to learn is through extension of the example base.

The translation templates generated by Güvenir and Cicekli (1998), for instance, differ according to the similarities and dissimilarities found in the reference text. Translation templates in this system thus reflect holistic aspects of the example translations. The way in which morphological analyses is processed is, however, independent from the translation examples and is thus a molecular aspect in the system.

Similarly, the ReVerb EBMT system (Collins, 1998) makes use of holistic components. The reference text is part-of-speech tagged. The length of translation segments as well as their most likely initial and final words are calculated based on proba-



•1: Sato and Nagao (1990)
•2: Carl (1999)
•3: Güvenir and Cicekli (1998)
•4: Zer (1997)
•5: Heyn (1996)
•6: Collins (1998)
•7: Brown (1997)
•8: Brown et al. (1990)
•9: McLean (1992)

Figure 2: A Model of Competence for CBMT

bilities found in the reference text.

4 Coarse vs. Fine Graining CBMT

One task that all MT systems perform is to segment the text to be translated into translation units which — to a certain extent — can be translated independently. The ways in which segmentation takes place and how the translated segments are joined together in the target language are different in each MT system.

In (Collins, 1998) segmentation takes place on a phrasal level. Due to the lack of a rich morphological representation, agreement cannot always be granted in the target language when translating single words from English to German. Reliable translation cannot be guaranteed when phrases in the target language - or parts of it - are moved from one position (e.g. the object position) into another one (e.g. a subject position).

In (Güvenir and Cicekli, 1998), this situation is even more problematic because there are no restrictions on possible fillers of translation template slots. Thus, a slot which has originally been filled with an object can, in the translation process, even accommodate an adverb or the subject.

SBMT approaches perform fine-grained segmentation. Brown et al. (1990) segment the input sentences into words where for each source-target language word pair translation probabilities, fertility probabilities, alignment probabilities etc. are computed. Coarse-grained segmentation are unrealistic because sequences of 3 or more words (so-called n-grams) occur very rarely for $n \geq 3$ even in huge learning corpora[1]. Statistical (and probabilistic) systems rely on word frequencies found in texts and usually cannot extrapolate from a very small number of word occurrences. A statistical language model assigns to each n-gram a probability which enables the system to generate the most likely target language strings.

[1] Brown et al. (1990) uses the Hansard French-English text containing several million words.

5 A Competence Model for CBMT

A competence model is presented as two independent parameters, i.e. *Coverage* and *Quality* (see Figure 2).

- **Coverage** of the system refers to the extent to which a variety of source language texts can be translated. A system has a high coverage if a great variety of texts can be translated. A low-coverage system can translate only restricted texts of a certain domain with limited terminology and linguistic structures.
- **Quality** refers to the degree to which an MT system produces successful translations. A system has a low quality if the produced translations are not even informative in the sense that a user cannot understand what the source text is about. A high quality MT-system produces user-oriented and correct translations with respect to text type, terminological preferences, personal style, etc.

An MT system with low coverage and low quality is completely uninteresting. Such a system comes close to a random number generator as it translates few texts in an unpredictable way.

An MT system with high coverage and "not-too-bad" quality can be useful in a Web-application where a great variety of texts are to be translated for occasional users which want to grasp the basic ideas of a foreign text. On the other hand a system with high quality and restricted coverage might be useful for in-house MT-applications or a controlled language.

An MT system with high coverage and high quality would translate any type of text to everyone's

satisfaction. However, as one can expect, such a system seems to be not feasible.

Boitet (1999) proposes "the (tentative) formula: $Coverage * Quality = K$ "where K depends on the MT technology and the amount of work encoded in the system. The question, then, is when is the maximum K possible and how much work do we want to invest for what purpose. Moreover a given K can mean high coverage and low quality, or it can mean the reverse.

The expected quality of a CBMT system increases when segmenting more coarsely the input text. Consequently, a low coverage must be expected due to the combinatorial explosion of the number of longer chunks. In order for a fine-graining system to generate at least informative translations, further knowledge resources need be considered. These knowledge resources may be either pre-defined and molecular or they can be derived from reference translations and holistic.

TMs focus on the quality of translations. Only large clusters of meaning entities are translated into the target language in the hope that such clusters will not interfere with the context from which they are taken. Broader coverage can be achieved through finer grained segmentation of the input into phrases or single terms. Systems which finely segment texts use rich representation languages in order to adapt the translation units to the target language context or, as in the case of SBMT systems, use holistic derived constraints.

What can be learned and what should be learned from the reference text, how to represent the inferred knowledge, how to combine it with pre-defined knowledge and the impact of different settings on the constant K in the formula of Boitet (1999) are all still open question for CBMT-design.

6 Conclusion

Machine Translation (MT) is a meaning preserving mapping from a source language text into a target language text. In order to enable a computer system to perform such a mapping, it is provided with a formalized theory of meaning.

Theories of meaning are characterized by three dichotomies: they can be *holistic* or *molecular*, *austere* or *rich* and they can be *fine-grained* or *coarse-grained*.

A number of CBMT systems - translation memories, example-based and statistical-based machine translation systems - are examined with respect to these dichotomies. In a system that uses a rich theory of meaning, complex representations are computed including morphological, syntactical and semantical representations, while with an austere theory the system relies on the mere graphemic surface form of the text. In a holistic implementation meaning descriptions are derived from reference translations while in a molecular approach the meaning descriptions are obtained from a finite set of predefined features. In a fine-grained theory, the minimal length of a translation unit is equivalent to a morpheme while in a coarse-grained theory this amounts to a morpheme cluster, a phrase or a sentence.

According to the implemented theory of meaning, one can expect to obtain high *quality* translations or a good *coverage* of the CBMT system.

The more the system makes use of coarse-grained translation units, the higher is the expected translation quality. The more the theory uses rich representations the more the system may achieve broad coverage. CBMT systems can be tuned to achieve either of the two goals.

References

Christian Boitet. 1999. A research perspective on how to democratize machine translation and translation aides aiming at high quality final output. In *MT-Summit '99*.

Peter F. Brown, J. Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, F. Jelinek, Mercer Robert L., and Roossin P.S. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16:79–85.

Ralf D. Brown. 1997. Automated Dictionary Extraction for "Knowledge-Free" Example-Based Translation. In *TMI-97*, pages 111–118.

Michael Carl. 1999. Inducing Translation Templates for Example-Based Machine Translation. In *MT-Summit VII*.

Bróna Collins. 1998. *Example-Based Machine Translation: An Adaptation-Guided Retrieval Approach*. Ph.D. thesis, Trinity College, Dublin.

Michael Dummett. 1975. What is a Theory of Meaning? In *Mind and Language*. Oxford University Press, Oxford.

Halil Altay Güvenir and Ilyas Cicekli. 1998. Learning Translation Templates from Examples. *Information Systems*, 23(6):353–363.

Matthias Heyn. 1996. Integrating machine translation into translation memory systems. In *European Asscociation for Machine Translation - Workshop Proceedings*, pages 111–123, ISSCO, Geneva.

Ian J. McLean. 1992. Example-Based Machine Translation using Connectionist Matching. In *TMI-92*.

Makoto Nagao. 1989. *Machine Translation How Far Can It Go*. Oxford University Press, Oxford.

S. Sato and M. Nagao. 1990. Towards memory-based translation. In *COLING-90*.

Zeres GmbH, Bochum, Germany, 1997. *ZERESTRANS Benutzerhandbuch*.