Towards a Model of Competence for Corpus-Based Machine Translation

Michael Carl
Human Language Technology Center
Hong Kong University of Science and Technology
Clear Water Bay, Kowloon, Hong Kong
lrac@cs.ust.hk

Abstract

A translation is a conversion from a source language into a target language preserving the meaning. A huge number of techniques and computational approaches have been experimented in order to translate natural languages automatically, yet no satisfactory solution has been found. This paper examines approaches to corpus-based machine translation (CBMT). In CBMT, a set of reference example translations is given to the MT system. These are analyzed and compiled into the system's internal representation according to the theory of meaning the system implements. The representations, then, serve as a basis to translate new sentences.

This paper discusses three main approaches in the CBMT paradigm: the memory-based approach (e.g. translation memories (TM)), the example-based approach (EBMT) and the statistical-based approach (SBMT). Concrete CBMT systems are discussed in light of the theory of meaning (preservation) they implement. This discussion, then leads to a model of competence for CBMT systems. The paper concludes that CBMT systems can be designed to achieve high *reliability* or broad *coverage*, though both seem to be mutually exclusive qualities.

1 Meaning Preservation in Machine Translation

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)). Therefore, an MT system which produces translations from an input text needs to be aware of the its meaning. The translation algorithm together with the data it uses code 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. The crucial question for MT is then: what formal components of a source language text contribute to its meaning that can equally be expressed in the target language? A number of research areas seek to approach — at least the first part of — the question. Cognitive science conceives humans as information processing devices; logics seek to uncover the logical structure inherent in texts and psycho-linguistics investigate how texts are perceived and produced by humans and what psychological devices are required to do so.

In linguistic research, texts are described on a number of levels and dimensions each contributing to its understanding and hence to its meaning. Having a perfect description of a text renders a possibility of a perfect translation. On the other hand, having a perfect machine translation system, would imply that one has a perfect theory of meaning. An "ideal" translation of a source language text is achieved when a target language text realizes equivalent features for all dimensions of the source language text. In order to describe texts, a number of levels of description are well established. Figure 1 shows a few dimensions which are frequently employed in linguistics to describe texts.

Although each dimension listed in Figure 1 is a field of research on its own, none of them is really independent from the others[1]. Synonyms, for instance, leave untouched the conceptual meaning of the text but may involve a change in morphology or syntax, when a word is replaced by its synonym. Also, the graphical layout of the text may affect other meaning components. In a user manual, different fonts may imply different understanding of the text passage. Titles in newspapers are usually highlighted and often differ syntactically and morphologically from the text body, etc.

Translation divergences are well studied in MT (cf. e.g. (Dorr, 1994)). They describe language specific phenomena which cannot be directly mapped into a target language. Such phenomena touch all levels of description and require a cross-dimensional mapping from the source language into the target language. For instance, some languages express some phenomena morphologically while others express the same phenomena syntactically. Also, thematic roles may change in the source and target language which may require a different morpho-syntactic realization of the target language. Besides this, concepts may be differently grained (e.g. finer) in the source and target languages.

Due to the cross-dependent mapping of description levels, one cannot usually expect to have an "ideal" translation which realizes equivalent features in all meaning dimensions. Depending on the context and the type of text, the equivalence of certain meaning dimensions appears to be more important than the equivalence of other meaning dimensions. It is, however, unlikely that all of them can be equally expressed in the target language so that the source and the target language text reach consistency. In some cases, one would have to abandon a lexical equivalence in favor of a conceptual equivalence as is the case when *validate* is translated into German *entwerten* (de-validate). In other cases, it might be required to find a phonetically equivalent form by ignoring, for instance, the lexical equivalence because rhyme and number of syllables are considered more important than other dimensions.

Even if we have a complete meaning description of the source language text in terms of the dimensions (e.g. as shown in Figure 1) we cannot be certain that an ideal translation can be generated in all cases. First, it might turn out that there are further meaning dimensions that we have forgotten, e.g. cultural background, intended readers, etc. but which appear to be important for translation. Second, in case there is no 1-to-1 mapping possible for all dimensions — which is more likely than not — we would have to decide which dimension to emphasize.

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. CBMT-systems assume that meaning equivalence holds for the reference example translations given to the system in a training phase. According to their intelligence, these systems try to figure out of what the meaning invariance consists in the

[1] I am grateful to Prof. Uszkoreit to whom I owe the original idea.

Figure 1: Dimensions of meaning equivalence in text translation

The source and the target language texts can be decomposed into a number of meaning dimensions. In (machine) translation one seeks to find target language expressions that reflect the meaning of the source language text. One cannot usually expect an equivalent mapping of all meaning dimensions.

Dimensions of Source language text			Dimensions of Target language text	
graphical	(*gra*)		(*gra*)	graphical
phonetical	(*pho*)		(*pho*)	phonetical
lexical	(*lex*)		(*lex*)	lexical
morphological	(*mor*)		(*mor*)	morphological
syntactical	(*syn*)		(*sym*)	syntactical
semantical	(*con*)	*translation* $\Longleftrightarrow$	(*sem*)	semantical
conceptual	(*con*)		(*con*)	conceptual
pragmatical	(*pra*)		(*pra*)	pragmatical
discourse	(*dis*)		(*dis*)	discourse
stylistic	(*sty*)		(*sty*)	stylistic
...	...		...	...

reference text and learn an appropriate source language/target language mapping mechanism.

A meaning equivalent target language translation can only be generated if an appropriate example translation is available in the reference corpus or an extracted mapping device here from. This is because the understanding of the new translation is derived from the understanding of the reference translation.

The reference corpus, thus, constitutes to the empirical basis of the translation learning process. The learning process' understanding capacities depend on the richness of the theory of understanding it implements. The more a system is aware of the different meaning dimensions inherent in the translation corpus, the finer its understanding capacities and the richer the representation will be.

An interesting question in CBMT systems is thus: what theory of meaning, or its components, 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 language 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 the 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 to all other expressions in that language. A molecular theory, therefore, gives 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 to the above 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 based on 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 rest of this paper is structured as follows. The next three sections discuss the dichotomies of theories of meaning, *rich vs. autere*, *molecular vs. holistic* and *coarse vs. fine grained* where concrete CBMT systems are related and classified according to the terminology. This discussion 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 *reliability*.

2 Rich vs. Austere CBMT

A common characteristic of all CBMT systems is that the understanding of meaning preserving translation mapping is derived from the understanding of the corpus of reference translation examples. The inferred translation mapping mechanism is used in the translation phase to generate new translations.

Collins (1999) distinguishes between Memory-Based MT, i.e. memory heavy, linguistic light and Example-Based MT i.e. memory light and linguistic heavy by the theories of meaning which they implement. 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 translation corpus. No abstraction or generalization of the reference examples takes place.

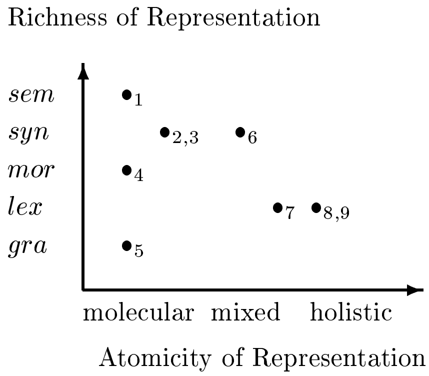
Translation Memories (TMs) are such purely memory based MT-systems. A TM e.g. TRADOS's Translator'sWorkbech (Heyn, 1996), and STAR's TRANSIT calculates the graphical similarity of the input text and the source side of the reference translations and return the target string of the most similar translation 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 in the reference corpus that contain a certain word.

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, then, 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 best 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, 1999; Güvenir and Cicekli, 1998; Carl, 1999)) are richer systems, where translation examples are stored

Figure 2: Atomicity and Richness of Representation

The graph depicts the richness and atomicity of meaning representation implemented in corpus-based machine translation systems. The atomicity of representation can be molecular, mixed or holistic. The richness can involve one or more levels of descriptions. In the graph, only the most abstract representations are plotted. As yet, no holistic/rich systems have been implemented.



- •$_1$: Sato and Nagao (1990)
- •$_2$: Carl (1999)
- •$_3$: Güvenir and Cicekli (1998)
- •$_4$: Zer (1997)
- •$_5$: Heyn (1996)
- •$_6$: Collins (1999)
- •$_7$: Brown (1997)
- •$_8$: Brown et al. (1990)
- •$_9$: McLean (1992)

as feature annotated and sometimes structured representations. Translation templates are generated which contain (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. Sentences are more finely decomposed into phrases and linguistic constituents e.g. NPs, PPs, subject, object, etc.

The ReVerb EBMT system (Collins, 1999) 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) (Turkish-English) 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 by Carl (1999) for the language pair English-German. 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. A thesaurus quantifies similarities of two or more concepts. 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.

Statistics based MT (SBMT) approaches implement austere theories of meaning. For instance, in Brown et al. (1993) 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 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 graphical symbols i.e. the ascii set. Richer systems use lexical, morphological, syntactical and/or semantical descriptions. The degree to which the set of descriptions is independent from the shape of the actual reference translations determines the molecularity of the implemented theory. The more the system internal representations are learned from the reference translations the more it becomes holistic.

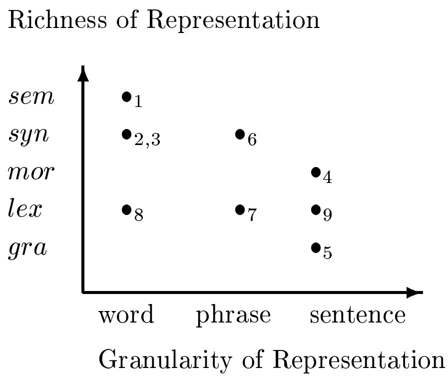
SBMT approaches e.g. (Brown et al., 1993) have a purely holistic view on languages. Every sentence of one language is considered to be a possible translation of any sentence in the other language. No account is given for the equivalence of the source language meaning and the target language meaning other than by means of global considerations concerning frequencies of occurrence in the reference corpus. 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 priori probability is derived from the reference corpus. 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 corpus 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 independant 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

Figure 3: Granularity and Richness of Representation

The graph puts into relation the granularity and the richness of representation. Finer grained representations allow a decomposition of the input text into shorter chunks. Richer representations code more information about the translation unit's properties. The coarser the granularity, the less rich a system may be.



$\bullet_1$: Sato and Nagao (1990)
$\bullet_2$: Carl (1999)
$\bullet_3$: Güvenir and Cicekli (1998)
$\bullet_4$: Zer (1997)
$\bullet_5$: Heyn (1996)
$\bullet_6$: Collins (1999)
$\bullet_7$: Brown (1997)
$\bullet_8$: Brown et al. (1990), Brown et al. (1993)
$\bullet_9$: McLean (1992)

of the graphic symbols nor does it modify the distance metric.
In more holistic implementations, the entities of descriptions change with the degree to which the learn texts differ.

The translation templates generated by Güvenir and Cicekli (1998), for instance, differ according to the similarities and dissimilarities found in the reference corpus. 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, 1999) makes use of holistic components. English-German examples are part-of-speech tagged and the length of translation segments as well as their most likely initial and final elements are calculated based on probabilities found in the reference text.

A still more holistic theory of meaning is realized in (Brown, 1996; Brown, 1997). Translation examples are retrieved from a database which have substrings with the input string in common. Adaptation of the target language is achieved by combining common substrings of the retrieved target language sentences by statistic means.

To conclude this section, the more a system makes use of molecular theories the more fragile it becomes. This is due to the inherent inconsistencies in natural language corpora from which representations are learned. Static similarity measures, as used in TMs and some EBMT systems, do not provide a general solution to this problem. They render a system more flexible for a domain they have been designed for. They cannot, however compete with the flexibility that holistic learning approaches bear.

4 Coarse vs. Fine graining CBMT

One task that all MT systems perform is to segment the text to be translated into translation units (TUs) which — to a certain extent — can be translated independently. This segmentation reflects the observation, that language itself and the meaning it transports often works in a compositional manner: blocks can be extracted from the text and understood independently from its context. Also, MT systems can be designed more easily if they make use of a number of small TUs which are may be independent from each other and reconstruct the target language text based on the TU translations. 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.

Once the TUs are translated into the target language, they need to be adapted to the target language context. Different MT approaches differ in the complexity of adaptations which they can perform and the number of adaptations which have to be performed. The need for adaptation arises when different TUs have to be combined to form a translation. For a correct adaptation, target language peculiarities such as case assignment, the choice of prepositional forms, case, number or person agreement, the choice of the part of speech etc. must be taken into account.

By chosing large TUs, the need for adaptations can be reduced to a minimum. TMs, for instance, coarsely decompose texts at a sentence level and fuzzy match it against examples contained in the database. Each sentence thus corresponds to a TU which is not further analyzed or decomposed into smaller constituents. The target language text is the mere concatenation of the TU's target language expressions without being further adapted or rearranged.

In rich systems, TUs contain knowledge which may stem from a lexical and/or a grammar database. This knowledge codes properties of the chunks, its behavior in the context and possible translation equivalences. From the added knowledge, the target language context can be rearranged and reconstructed. Therefore, rich systems may decompose the text into small segments. The granularity of the chunks, thus, depends on the adaptability of the MT system. The adaptability, in turn, depends on the (linguistic) richness of the internal representation. In MT systems which lack such rich internal representation, this ability is limited. On the other hand, the more a system has to adapt, the more it is prone to errors.

In (Collins, 1999) for instance, agreement cannot always be reconstructed in the target language, when translating single words from English to German which are part of larger constituents. Reliable translation cannot be guaranteed when phrases in the target language (or parts of it) are moved from one functional position into another one.

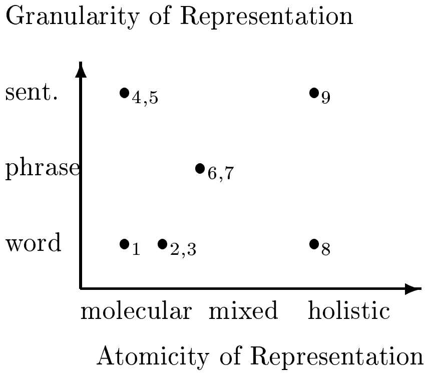
In (Güvenir and Cicekli, 1998), this situation is even worse because there are no restrictions on possible fillers of translation templation slots. Thus, a slot which has originally been filled with an object can, in the translation process, even accommodate an adverb or the subject.

SBMT approaches perform fine grained decomposition. Brown et al. (1990) decomposes the input sentences on a word level where for each source-target language word pair translation probabilties, fertility probabilities, alignment probabilities etc. are computed. Coarse grained decomposition cannot take place because sequences of 3 or more words (so-called n-grams) occur very rarely for $n > 3$ even in huge learning corpora[2]. Statistical (and probabilistic) systems rely on word frequencies found in the corpora and cannot usually extrapolate from very small

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

Figure 4: Atomicity and Granularity of Representation

The graph depicts system implementations with respect to their atomicity and granularity of representation. Almost any combination of parameters can be found in CBMT-systems.



$\bullet_1$: Sato and Nagao (1990)
$\bullet_2$: Carl (1999)
$\bullet_3$: Güvenir and Cicekli (1998)
$\bullet_4$: Zer (1997)
$\bullet_5$: Heyn (1996)
$\bullet_6$: Collins (1999)
$\bullet_7$: Brown (1997)
$\bullet_8$: Brown et al. (1990)
$\bullet_9$: McLean (1992)

numbers. A statistical language model assigns to each n-gram a probability which enables the system to generate the most likely target language strings.

Figures 2, 3 and 4 put the discussed parameters of a theory of meaning in a graphical relation. It can be seen that almost any combination of the parameters is a possible option for the design of a CBMT system. Which of the parameters are to be chosen in order to achieve the desired goals shall be discussed in the next section.

5 A Competence Model for CBMT

In the previous section the components of a theory of meaning are discussed as they are implemented in various CBMT systems. A number of examples are given and system architectures are presented. In this section, the expected competence of the different CBMT systems shall be examined in light of the parameters of the theory of meaning.

The competence model is presented as two independent parameters, *Coverage* and *Reliability* (see Figure 5). An MT system can either be designed to reproduce for a small language segment i.e. a sub-language or a controlled language with high fidelity and precision or it may be designed to perform informative, general purpose translations. In the former case, the system will have high reliability, whereas in the latter case, its coverage will be high. However, both properties are, to a certain extent, mutually exclusive.

- **Coverage** refers to the extent to which a great variety of source language texts can successfully be translated into the target language. A successful translation can be described as to be informative in the sense that allows a user to understand more or less the content of the source text.
- **Reliability** refers to the extent to which an MT system approaches an "ideal" translation (of a restricted domain) for a given purpose or for a given user. A reliable translation is user-oriented and correct with respect to text type, terminological preferences, personal style, etc.

TMs focus on the reliability 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. Reconstruction of the target language meaning is, thus, a simple concatenation of concepts.

TMs covers an input text to the extent that they can retrieve for the segmented input text at least one reference translation above a given threshold. TMs, thus, tend to have a small coverage because large TUs are unlikely to fit into different contexts of different text types. Due to the lack of representational richness, no transformation or adaptation of the retrieved target language sentence into the target language context takes place. TMs have no device to adapt their retrieval outcome based on former experiences. However, due to the size of the TUs, adaptation is not always required in the translated target language.

Broader coverage can be achieved when decomposing sentences into shorter TUs such as phrases or single terms. This is because the smaller the TUs are the more they are likely to fit into different contexts.

The adaptation mechanism which recomposes the target language words into a sentence must be sufficiently powerful to perform this task. Systems which decompose texts into fine segments use rich representation languages in order to adapt the TUs' translation into the target language context. In Sato and Nagao (1990), sentences are chunked into single words such that almost each word corresponds to a proper chunk. The algorithm prefers the substitutions of semantically similar word tokens in target language derivation trees. Collins (1999) performs a coarser chunking and replaces thematic related roles as fillers in translation templates. (Carl, 1999) adapts the part-of-speech and a set of external constraints of the templation slots and their fillers.

However, reliability of EBMT systems may suffer in case of a lack of sufficiently rich representation. In (Collins, 1999) for instance, agreement cannot always be reconstructed in the target language, when translating single words from English to German, and a reliable translation cannot be guaranteed when phrases in the target language (or parts of it) are moved from one functional position into another.

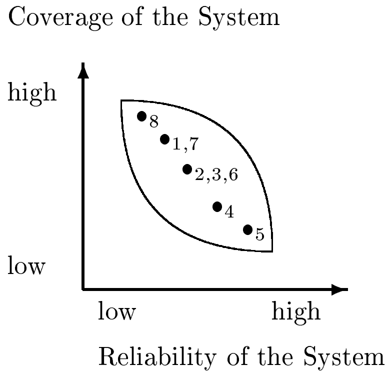
In (Güvenir and Cicekli, 1998), this situation is even worse because there are no restrictions on possible fillers of translation templation slots. Thus, a slot which has originally been filled with an object can, in the translation process even accomodate an adverb or the subject.

SBMT systems focus mainly on coverage. In contrast to TMs and EBMT systems, SBMT performs abstraction of the reference corpus[3]. Random variables are instantiated in the learning phase of the system and kept in memory as the basis for translation. When translating a new sentence, only values of the instantiated random variables are considered and the original reference corpus is no longer regarded as the guide translation. In this sense, SBMT systems do not seek to reconstruct the reference example

[3] According to Daelemans (1998) a distinction can be made between systems which generalize and systems which perform abstractions. When generalizing, the original data is kept in the system and further data-driven generalizations are added. In abstracting systems, the reference corpus only serves as a basis on which internal representations are created. Abstracting systems systems do not consider the original data in the working phase but rather only rely on the generated abstractions.

Figure 5: A Model of Competence for CBMT

The graph depicts the expected competence of the discussed CBMT systems. A translation system can either be tuned to achieve broad coverage or it can be designed to have high reliability. Both parameters seem to be mutually exclusive. The oval frame outlines the space of possible CBMT systems.



●$_1$: Sato and Nagao (1990)
●$_2$: Carl (1999)
●$_3$: Güvenir and Cicekli (1998)
●$_4$: Zer (1997)
●$_5$: Heyn (1996)
●$_6$: Collins (1999)
●$_7$: Brown (1997)
●$_8$: Brown et al. (1990)
●$_9$: McLean (1992)

translations with high precision but try to find the most likely translation sequences. Coverage in a statistics based MT system is achieved to the degree to which entries in the probability tables are available for the currently processed items.

The expected reliability of the system increases when chosing a coarser decomposition of 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 system to generate at least informative translations, and thus to accord to the coverage criteria when chosing fine grained decomposition, further knowledge resources need be considered. These knowledge resources may be either molecular and rich or they may be holistic and austere. Whereas exclusively holistic representations provide high coverage, molecular systems enable reliable translations.

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 must be provided with a formalized theory of meaning. This paper discusses components of a theory of meaning and investigates how these components are realized in different Corpus-Based MT (CBMT) systems.

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

A number of CBMT implementations, translation memories, example-based and statistical-based machine translation systems are examined with respect to these dichotomies. In all of the CBMT approaches, the following can be observed: A source language text is decomposed into a set of meaning units and each of which is, to a certain extent, independently translated into the target language. The set of translations is then recomposed into a target language text.

This paper then discusses parameters of a competence model for CBMT-systems. According to the shape of the implemented theory of meaning, one can expect to obtain *reliable* translations or a good *coverage* of the CBMT system.

The more the theory becomes holistic and/or rich, the more the system is designed to achieve broad coverage. The more the system makes use of coarse grained meaning units, the higher is the expected reliability. CBMT systems can be tuned to achieve either of the two goals. They may achieve a high coverage or they can be tuned to produce reliable translations, which respect to text type, terminological preferences, personal style etc.

References

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

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical Machine Translation: Parameter estimation. *Computational Linguistics*, 32:263–311.

D. Ralf Brown. 1996. Example-Based Machine Translation in the Pangloss System. In *COLING-96*.

D. Ralf 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. 1999. *Example-Based Machine Translation: An Adaptation-Guided Retrieval Approach.* Ph.D. thesis, Trinity College, Dublin.

Walter Daelemans. 1998. Abstraction is Harmful in Language Learning. In *Proceedings of NeMLaP*, pages 1–2, Sydney, January.

Bonnie Jean Dorr. 1994. Machine Translation divergences. *Computational Linguistics*, 20(4):597–633.

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 Exaples. *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.*