Semantics in Machine Translation

N. Klyueva
Charles University Prague, Faculty of Mathematics and Physics, Prague, Czech Republic.

Abstract. In this paper we introduce Machine Translation methods, their main paradigms and architectures. The importance of semantic information in MT systems is shown and the problem of ambiguity is discussed. We look at the previous MT systems of our interest and the usage of semantics in them. The notion of semantic features and their application in the systems is therefore presented.

1 Introduction

Herein, we briefly review Machine Translation (MT) techniques, such as MT paradigms and architectures and discuss the problem of semantics and ambiguity in MT. We describe the approach that will better suit our goal to make the Czech-to-Russian MT system. The Russian and Czech languages belong to the same group, and they have many features in common, so we have chosen the solution, that was proposed by Hajič [Hajič et al., 2003], the Rule-Based MT system.

Considering that there was an attempt to build the MT between this pair [Oliva, 1989], we are going to make use of the linguistic data, that were gathered during the experiment. The data mentioned is the annotated Czech-to-Russian dictionary, which we are going to include into the new module. The valency information can be extracted from the dictionary and partly used in our first experiments. What's more, the experience of using the semantic features can help our future work with the Word Sense Disambiguation problem.

2 Machine Translation in General

The MT systems can be classified according to various criteria. First, we can compare the systems with respect to the information that they are based on. The oldest systems were **Rule-Based**, they contained manually annotated dictionaries and manually written rules of transfer (METEO, SYSTRAN). They were labor-intensive and sometimes it took years to make a system.

Data-Driven MT. This approach appeared in the mid 90s. It used huge parallel corpora and n-gram statistic model to search for the most probable translation equivalent. The first system was developed by IBM, but now Google model is the best one, exploiting the enormous amount of data.

Still, all of the previous approaches had their disadvantages, so nowadays the new **Hybrid MT** systems are constructed (Euromatrix).

MT systems can also be classified on the basis of their design. There are at least three architectures: direct, transfer and interlingua. These levels are generally introduced in the Machine Translation triangle:

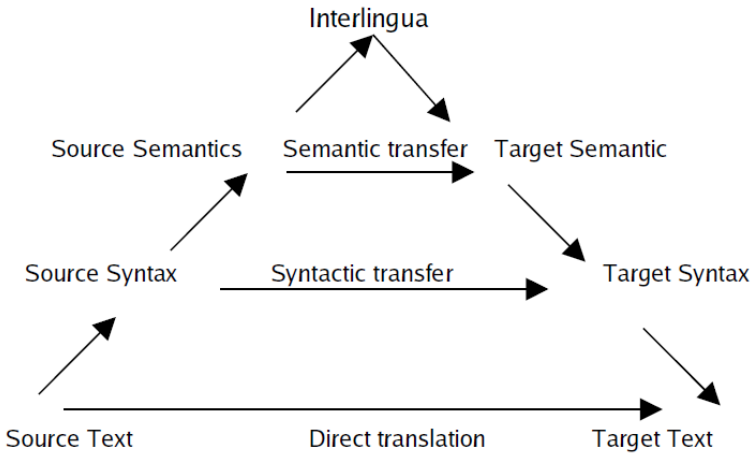

Figure 1. MT Triangle.

Direct systems provide word-by-word translation from a source to target language. In this case, the two languages should be closely related. For instance, the direct architecture was used in the translation between Czech and Slovak languages [Hajič et al., 2003]. Next languages that were added to Česílko were Polish and Lithuanian. Additional rules for syntactic transfer were added into the system, because the syntax of the source and target languages is not similar. The most vivid example is adjective postposition in Polish in contrast to Czech. The experiment with the translation from Czech to Lithuanian has showed that a deeper annotation should be made for the languages with different syntactic structure a.

Transfer translation. The source text first goes through tagger, where it receives the appropriate morphological features (part of speech, tense, case, voice etc.). The morphological information is then analyzed by a syntactic parser. The output of a syntactic analysis is a syntactic tree, with the nodes having morphological information about the word. The advantage of a transfer system is that the modules (syntactic, semantic synthesis and analysis) are language independent, and they can be reused in translation into other languages.

The Machine Translation systems that go deep into language meaning are called **Interlingua** systems. In Interlingua systems no transfer is needed, only synthesis and analysis is made. Here researchers face the problem of Universal language (UNL) – one, which encode all possible meanings in every natural language. The system will be based on one formal representation, and each language will have a stage of synthesis and analysis connected to UNL. The disadvantage of the architecture is that it will be hard to specify language information – there will be no lexical items, only universal semantic primitives.

3 MT systems of our interest

Our first experiments focus on the Czech and Russian language pair. The history of the MT system between these two languages is a long one. The first project, Ruslan, started in the late 1980s. **Ruslan** was a *rule-based MT* system from Czech into Russian, with morphology, syntax and a bit of semantic information stored in the dictionary. We are going extract the useful information from the dictionary and use it on the transfer level.

As the basis of the system we are going to take a more modern application – the system Česílko mentioned in the previous section as the direct MT system. Consider the difference between two dictionary entries in these MT systems:

Table 1. Differences between Ruslan and Česílko

Ruslan's entry	***Česílko's entry***
NA2PAD==H(@(*A),FI1023,**IDEJA)**.	translate('**nápad**', _, '**ideja**', [], []).
SNADN==M(MI2884,**LEGKIJ)**.	translate('**snadný**', _,'**legkij**', [], []).
BRZY == 6($(**SKORO)**) .	translate('**brzy**', _,'**skoro**', [], []).
KOMPILOVA==R(5,K,?(N(N,A(**KOMPILJACIJU**(G)),I(_SE),I),NOPAS,23,**OSUS4ESTVLJAT6**)	translate('**kompilovat**',_,'**kompilirovat1**', [], [], [[ag, subj],[pat, obj]]).

Our aim is to extract data from the first system and transfer it into the second formalism. The main difference is that in Ruslan we have word roots in Czech and the conjugation/ model (the symbol that precedes the left bracket – H,M,6). So we will first decode the symbols and transfer them into the lemmas of Česílko. Though, we should count with the problem of language development. While in the 80s the word *kompilovat* was translated as a collocation *osuščestvljat'c kompiljaciju*, in modern Russian it is translated just as *kompilirovat'*.

The next type of information, that we are going to extract, is valency information. Though new and more user-friendly systems of representing valency frames were built recently, we will start our research with Ruslan frame, as only there a mutual information on the two languages can be found:

NECHA==R(5,D,?(N(N),A(A),/,N(N),I(I)),26,OSTAVIT6).

translate('nechat', _,'ostavit'', [], [], [[ag,sub], [pat,obj]]).

There is one more type of semantic information coded in Ruslan's entry – semantic features. In the next chapter we discuss the problem of the semantic annotation and why it is so crucial for the process of MT.

4 Ambiguity in MT

4.1 Word Sense Disambiguation

The key problem that we come across in language processing is an ambiguity on different language levels: syntactic, lexical, and semantic. Ambiguity is a feature of natural languages to produce two and more meanings of a word. For text disambiguation we generally use semantic annotation. The MT system faces several types of natural language ambiguity:

Syntactic ambiguity: *A dog saw a cat with a telescope.*

Lexical ambiguity: *Book* (En) can be translated into Czech as either *kniha* or *reservovat*

Semantic ambiguity:

homography (bank, bug)

polysemy ((Ru)*obucheniye*) can be translated as *1) teaching* or *2) learning*)

There are different approaches for the problem. The most natural way of word sense disambiguation (WSD) is to look at the context of the word, and chose the most appropriate variant. Rough disambiguation can be made on the basis of the keyword of the document. It is more probable to choose bank as river bank meaning in geographical contexts and bank as a financial institution in the context of economic studies. Nowadays with regard to statistic methods prevailing in linguistics, the WSD is exploiting the corpus-based approach.

4.2 Semantic features

One of the oldest methods of semantic annotation is assigning the semantic features to words. Semantic features can be defined as elementary components of the meaning, so they can be found in the dictionary definition of a word. Here is the example of how two words are distinguished by the semantic features: Man [+HUMAN], [+MALE], [+ADULT] vs. Woman [+HUMAN], [-MALE], [+ADULT].

Semantic features in Ruslan were incorporated in the dictionary. The following example shows how they help to disambiguate the word:

DLOUH==M(RS(+(*INT)), MI2289, DLINNYJ)

DLOUH==M(RS(-(*INT)), MI2276, DOLGIJ)

The ambiguity here appears in a target language: while in Czech one can say *Dlouhá doba* a *Dlouhá ulice*, in Russian the distinction between time and space features is made (*долгое время*, *длинная улица*). The semantic features +(*INT), -(*INT) help to disambiguate two senses of one word. It took a great effort to annotate data, almost ten hundred lexical entries, and still the desirable result was not achieved. Further investigations showed, that in many cases semantic features would make the translation even worse. For example, a semantic feature *A (Animated) was assigned to the subject of the verb „běhat". Nevertheless, there are contexts, such as „program is running", where the subject is not animated, so the system will fail to make the appropriate connection. [Kuboň, 2001]

Our task in the area of semantic annotation will be to create the semantic annotation tool for the Czech-to-Russian MT. In the experiment we are going to use the experience from the previous systems and such semantic resources, as Czech and Russian WordNet, Vallex for valency information and Pustejovsky's inheritance structure. Collecting the set of appropriate semantic features for the two languages and extracting valency information will be the start point for our investigation into semantics in Machine Translation.

5 Conclusion and Future work

In this paper we have presented some aspect of the Machine Translation and Word Sense Disambiguation. We described the data that will help us to create a new Czech-Russian module within the Česílko MT project. The problem of the semantic annotation for the MT systems between the closely-related languages is an interesting theoretical question for future study.

References

Dorr, B.J. et. al. : A Survey of Current Paradigms in Machine Translation, In the Advances in Computers, Vol. 49, M. Zelkowitz (Ed.), Academic Press, London, 1999.

Hajič, J., Homola, P. and Kuboň, V. : A simple multilingual machine translation system. In *Proceedings of the MT Summit IX*, New Orleans, 2003.

Kuboň, V. : Problems of Robust Parsing of Czech. Ph.D. thesis, MFF UK Prague, 2001.

Lopatková, M.: Valency in the Prague Dependency Treebank: Creating the Valency Lexicon. In *Prague Bulletin of Mathematical Linguistics*, 79–80, MFF UK Prague, 2003.

Oliva, K.: A Parser for Czech Implemented in Systems Q, in Explizite Beschreibung der Sprache und automatische Textbearbeitung, MFF UK Prague, 1989.