METHOD FOR AN AUTOMATIC GENERATION OF A SEMANTIC-LEVEL CONTEXTUAL TRANSLATIONAL DICTIONARY

Dmitry Kan

Department of Technology of Programming, Saint-Petersburg State University, Universitetsky prosp. 35, Peterhof, Russia

Keywords: Translational Dictionary, Semantic Analyzer, Computer Semantics, Machine Translation, Disambiguation.

Abstract: In this paper we demonstrate the semantic feature machine translation (MT) system as a combination of two fundamental approaches, where the rule-based side is supported by the functional model of the Russian language and the statistical side utilizes statistical word alignment. The MT system relies on a semantic-level contextual translational dictionary as its key component. We will present the method for an automatic generation of the dictionary where disambiguation is done on a semantic level.

1 INTRODUCTION

There are two fundamental approaches to Machine Translation (MT): rule-based approach and statistical approach, which is based on streams of input data. Each of these two approaches has their advantages: the rule-based approach deeply studies and formalizes the linguistic rules of a natural language; statistical approach gives an opportunity to rapidly prototype new algorithms with application to several different natural languages using data streams. Along with it, there are as well disadvantages attributed to each of the two fundamental approaches to MT: rule-based approach commonly lacks automation of language formalization process and is usually bound to one natural language (or very few similar languages); statistical approach generally avoids deep view into the properties of a natural language giving away the task of language formalization to a numerical algorithm. In this paper we would like to present an ongoing project of an MT system, that merges rule-based and statistic approaches in its components where it is possible.

Since the main component of the system is translational dictionary, we prepare the grounds of its automatic generation in Section 2. The rule-based side of the system is supported by the functional model of Russian language in Tuzov, 2004. It is described in brief in Section 3. We utilize the results of Sections 2 and 3 for automatic creation of a semantic-level contextual translational dictionary in Section 4. Finally, we present the experimental MT system in Section 5. We list the main features of the presented MT system in Section 6.

Classic MT triangle (cf. Fig. 1) separates semantic and syntactic transfers.

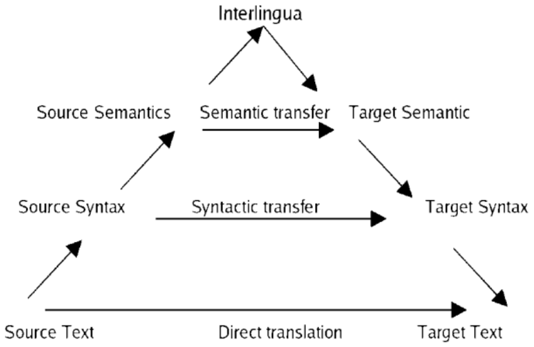

Figure 1: MT triangle adopted from Klueva, 2007.

The functional theory of the Russian language Tuzov, 2004 however shows that these two levels are interconnected. A morphological surface may point to two or more parts of speech. All of them can be equally considered as candidates on the syntactic level. However a word's semantics is required for a successful final resolution of a sentence meaning (see Section 3 for more detail).

Bennett, 1990 questions the need of a full-blown semantic analysis for MT and instead suggests advancing the „semantic feature system‟ to achieve

Kan D..
METHOD FOR AN AUTOMATIC GENERATION OF A SEMANTIC-LEVEL CONTEXTUAL TRANSLATIONAL DICTIONARY.
DOI: 10.5220/0003466504150418
In *Proceedings of the 6th International Conference on Software and Database Technologies* (ICSOFT-2011), pages 415-418
ISBN: 978-989-8425-77-5

cost effectiveness. In this paper we present the MT system, which has semantics coded on the level of dictionary entries and uses semantic analyzer to represent the meaning of an input sentence.

The related work Homola, 2009 shows how to build a translational dictionary between Chech and English with the statistical word alignment. However Homola, 2009 does not provide a method of resolving the word ambiguities. The main challenge during generation of a translation dictionary is resolving ambiguity of numerous word sequence pairs. In this work we suggest an approach which allowed us to solve the task by semantic interpretation of each dictionary entry.

2 TASK OF WORD ALIGNMENT

To build an MT system which uses statistic modelling of a natural language one needs a parallel corpus. Based on the corpus a model of translation is built and it contains phrase translational dictionary. In the process of constructing the dictionary, phrases of the parallel corpus get mapped together through maximizing the probability of their co-occurrence in the parallel corpus. Maximization is conducted over all possible word sequences of two aligned sentences in two languages (cf. Fig. 2).

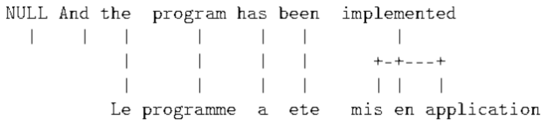
NULL And the program has been implemented
Le programme a ete mis en application

Figure 2: Example of one word-level alignment of English and French sentences adopted from Och, 1999.

The algorithm of word alignment is described in Al-Onaizan et al, 1999 and Och, 1999, while the full mathematical formulation of statistical MT is given in Brown, Della Pietra, Della Pietra and Mercer, 1993. The toolset Moses by Koehn, 2007 allows building a statistical MT system and includes GIZA++ as one of its component. GIZA++ implements the above algorithms for word alignment and outputs the following structure for each pair of sentences in the parallel corpus:

```
Desperate to hold onto power , Pervez
Musharraf has
discarded Pakistan ' s constitutional
framework and
declared a state of emergency .
NULL ({20}) В ({})
отчаянном ({1 3 4})
стремлении ({2}) удержать ({}) власть
({5}) ,
({6}) Первез ({7}) Мушарраф ({8})
отверг ({9 10})
конституционную ({14 15})
систему ({})
Пакистана ({11 12 13}) и ({16})
объявил ({17}) о ({18})
введении ({})
чрезвычайного ({19 21})
положения ({}) . ({22})
```

The above structure represents a mapping of words in Russian sentence into sequences of words in its English translation. Table 1 contains the mapping for the above example.

Table 1: Word alignment for English and Russian sentences.

Russian	English
NULL	of
отчаянном	Desperate to hold
стремлении	to
власть	power
,	,
Первез	Pervez
Мушарраф	Musharraf
отверг	has discarded
конституционную	constitutional framework
Пакистана	Pakistan ´ s
и	and
объявил	declared
о	a
чрезвычайного	state emergency
.	.

3 COMPUTER SEMANTICS

According to the theory in Tuzov, 2004 any natural language is functional in strict mathematical sense. Each sentence can be represented in a form of superposition of its functions-words:

$$S = F(f_1(w_{11},...,w_{1k}),...,f_n(w_{n1},...,w_{nl})), \\ w_{ij} \neq w_{hm}, \forall i \neq h, j \neq m \tag{1}$$

Definitional domain and values domain of $F,f_1,..,f_n$ belong to reality. In (1) word inequalities mean, that we count each word only once. This holds even if there are several repetitions of a word with the same or different semantics in the input sentence. In the process of a sentence analysis, semantic analyzer implemented by Tuzov, 2004 operates with the word senses extracted both from the hierarchical ontology with more than 2000 classes and from the syntactic-semantic dictionary. Since a dictionary word is generally represented as *n*-ary function with semantic constraints, its final semantics in the sentence depends on the exact words in its context within the sentence. The Tuzov's syntactic-semantic dictionary contains more than 150, 000 semantic formulas.

4 TRANSLATIONAL DICTIONARY

For creation of the MT system that is based on the functional theory in Tuzov, 2004 we need to translate the semantic dictionary onto the target language. For the process automation we have chosen the method described in the Section 2. For the parallel corpus we have used list of parallel sentences in Russian and English from the package UMC Klyueva and Bojar, 2008. GIZA++ has generated 1,3 million of phrase pairs, including duplicates. Applying the semantic analyzer on each Russian sentence we have obtained the semantic alternatives of each of the words in the built pairs, which correspond to their local context. As a result, each word in the original translational phrase dictionary was substituted with its semantic formula. This solves the disambiguation on semantic level. After removing the duplicates from the modified dictionary, the final version of semantic-level contextual translational dictionary has been built with about 18, 000 word pairs. The dictionary is subject to further clean up procedures and enrichment. Here is an extract from the final dictionary:

```
В Y1>HabU(Y1:,ПРЕД:Z1) \\ <149>--->Within
В Y1>Loc(Y1:,ВНУТРИ$12/313/05(ПРЕД:Z1)) \\
<146>--->at
В Y1>Loc(Y1:,Oper01(#,ПРЕД:Z1)) \\ <208>---
>In
В Y1>Loc(Y1:,ПРЕД:Z1) \\ <224>--->Throughout
...
НА Y1>Direkt(Y1:,ВЕРХ$12/141/05(ВИН:Z1)) \\
<67>--->at
НА Y1>Direkt(Y1:,РОД:Z1) \\ <100>--->on
НА Y1>Direkt(Y1:,РОД:Z1) \\ <69>--->for
...
ОБРАЗ (РОД:Z1) \\ <2>--->a way
ОБЩЕМИРОВОЙ
A1>Rel(A1:НЕЧТО$1,ПОЛНЫЙ$12/207/05(МИР$1227)
)
\\ <1>--->global
...
```

Each dictionary entry contains semantic formula corresponding to the original Russian word and its English analogue. One important property of the dictionary is that its entries are context dependent. This is provided by two circumstances: 1) each sentence in Russian had its expert translation into English and 2) each Russian word has been attributed with a semantic formula that was a result of semantic assembling of the corresponding Russian sentence.

Consider one entry of the above extract in more detail:

```
    В Y1>HabU(Y1:,ПРЕД:Z1) \\
    <149>--->Within
```

The semantic formula on the left of ---> sign has several components: the word "*В*" (the Russian preposition with a lot of meanings, roughly corresponding to the English prepositions *in, at, within, into, to, of* etc); the basis function *HabU(x,y)* with its arguments (the function defines that *x* possesses *y*), which in the case of *HabU* are Y_1 and Z_1 (prepositional case); \\ sign followed by the order number of the semantic alternative in the semantic-syntactic dictionary.

Another example:

```
    Y1>Loc(Y1:,ВНУТРИ$12/313/05(ПРЕД:Z1))
\\ <146>--->at
```

The second argument of *Loc(x,y)* basis function (defines, that *x* is located in / at *y*) is itself a function-preposition that takes one argument in prepositional case. The second argument has the name "*ВНУТРИ*" which is appended with ontological class number $12/313/05. In this class number, 12 refers to physical objects, 313 refers to inhabited locality, 05 refers to physical position of an object in prepositional case which is expressed as adverb in Russian ("ВНУТРИ" is both "where?" and "how?").

5 EXPERIMENTAL MT SYSTEM

The semantic-level translational dictionary obtained in the Section 4 forms the ground for the experimental Russian to English MT system. In order to achieve fluency on the target language side and to reduce the noise in the automatically generated semantic-level translation dictionary, we have devised the Semantic Machine Translation Model (*SMTM*) for translating sentence *P* onto the target language L_2:

$$SMTM_P = \arg\max_{i=\overline{1,n}} \Omega_i^S(t_1,\ldots,t_m) = \arg\max_{\substack{k=\overline{1,m-1}\\ l=\overline{2,m}}} \sum_i \delta_i^s(t_k,t_l), \tag{2}$$

where:

$$\delta_i^S(t_k,t_l) = \begin{cases} 1, & t_k t_l \in L_2 M \\ 0, & t_k t_l \notin L_2 M \end{cases} \tag{3}$$

Index S in (2), (3) suggests that the definitional domain of functions δ_i^S and Ω_i^S is the set of semantic formulas coded with symbols $t_l \in L_2$ in the translational dictionary. L_2M in (2) is the statistical model of L_2. The translation algorithm starts with translation of an input Russian sentence semantic representation. It then extracts the calculated semantic alternatives for each of the words and maps them onto their English translations. Form the final sentence in English using the model (2). Some of the translations show the advantage of semantic processing over statistical modelling Moses by Koehn, 2007 in that the system picks the correct semantic alternatives of polysemic words and their correct translations. The experimental MT system also translates more words than Moses from Russian to English, because it operates with the word lemmas while Moses is sensible to word surfaces.

6 FEATURES OF THE EXPERIMENTAL MT SYSTEM

The presented MT system is in its early stage of incorporating the semantic component into the process of MT. We have utilized statistical methods for automating the construction of a semantic-level translational dictionary and functional theory of natural language to resolve the ambiguity and introduce semantic context. Here is the list of the main features of the MT system:

- Dictionary entries contain semantic attributes of the Russian words;
- Each entry represents a sample of a context extracted using statistical word alignment and coded with the corresponding semantic formula;
- The MT system is automatically extendable through acquiring new parallel corpora and applying the method described in Section 4.

We have presented the MT system that can be categorized as a semantic feature system according to Bennett, 1990, which has complex semantic analysis system under the hood.

REFERENCES

Al-Onaizan Y., Curin J., Jahr M., Knight K., Lafferty J., Melamed D., Och F. J., Purdy J., Smith N. A., Yarowsky D. 1999. *Statistical Machine Translation. Final report.* Statistical Machine Translation, JHU Workshop.

Bennett W. S. 1990. *How Much Semantics is Necessary for MT systems?* Proceedings of the third International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages, Linguistic Research Center, University of Texas, Austin, TX:261-270.

Brown P. F., Della Pietra V. J., Della Pietra S. A., Mercer R.L. 1993. *The Mathematics of Statistical Machine Translation: Parameter Estimation.* Computational linguistics, 19(2):263-311.

Klueva N., Bojar O. 2008. *UMC 0.1: Chech-Russian-English Multilingual Corpus*. Proceedings of the conference “Corpora 2008”, Saint-Petersburg.

Klueva N. 2007. *Semantics in Machine Translation.* WDS’07 Proceedings of Contributed Papers, Part I:141-144.

Koehn P. et al. 2007. *Moses: Open Source Toolkit for Statistical Machine Translation*. Annual meeting of the Association for Computational Linguistics, demonstration session, Prague, Chech Re-public.

Och F. J. 1999. *An Efficient Method for Determining Bilingual Word Classes.* Ninth conf. of the Europ. Chapter of the Association for Computational Linguistics, EACL’99:71-76.

Tuzov V., 2004. *Computer Semantics of the Russian Language,* Saint-Petersburg State University Publishing House. Saint-Petersburg.