A Possible Solution to the Problem of Machine Translation of Verb Forms from Bulgarian to English

Todor Lazarov
Department of computational linguistics, IBL-BAS
todorlazarov91@abv.bg

Abstract

The paper`s main subject is concerned with the problems related to machine translation of verb forms from Bulgarian to English. In separate sections of this article we discuss the problems related to differences between word formation in both languages and differences in the information that the verb forms grammaticalize. We also introduce the idea of implementing the statistical method of machine translation altogether with the rule-based method as a proposal for future research and the possible practical and theoretical outcomes.

1. Introduction

The verb is a part of speech, which denotes changeable in time actions and states of objects, i.e. dynamic properties. From what has been said, it follows that one of the essential features of the verbs is to give information about the relation of these properties with other elements on the temporal axis, which has an absolute referential point – the act of speaking (now). The grammatical tense serves for expressing the different types of correlation of events and actions in time. Languages differ in the number and the types of tenses that they can express. The two languages, which this article discusses, share several common features, but main object of interest for us are the numerous distinctive features of the Bulgarian verb system, which make it interesting and difficult for formal description for the purposes of machine translation.

Translating verb forms is very difficult even for human translation – even though the verb systems of both English and Bulgarian share numerous common characteristics, they differ in the manner in which they express the relations between events and points on the temporal axis, the action denoted by the verb and the information about these events. Nevertheless, as we speak about the opportunities of machine translation, both languages are resource rich, which makes theoretical and practical researches about different aspects of them reliable and the gathered data – practical for the purposes of natural language processing and machine translation. In this paper we propose a hypothesis that implementing statistical language modelling with rule-based machine translation can improve our knowledge not only about the relation of the verb systems of Bulgarian and English, but also about the structural dependencies between these two languages. In future we can use the results of this research for the purposes of achieving higher quality of translation and better understanding of both languages.

2. Main differences and similarities between the verb systems of both languages.

A well-known fact is that one of the most distinguishing properties of the Bulgarian language is its well-developed verb system. On one hand, regarding the semantic of the verb forms, the Bulgarian verb can have over 2000 forms with different grammatical meaning. The literature on Bulgarian tense system consists of many disagreements, mainly about the grammatical categories that it contains and the differential properties of these categories. In this paper we acknowledge the view of the Bulgarian tense category as a hyper-category, in which the meaning is formed by the relationship between the individual members of different categories, as it is possible to include additional elements that can

modify the meaning. (Gerdzhikov, 2000) The English temporal system share the same feature – it can be considered as a hyper-category, but a main difference regarding the Bulgarian temporal system is the much smaller number of grammaticalized meanings in English. Other distinctive feature of the Bulgarian verb is the potential to grammaticalize the indication of the nature of evidence for a given statement – the category of evidentiality (Nitsolova, 2008: 261). These characteristics contribute for the difficulties in the process of semantic transfer during translation.

On the other hand, the word formation of the verbs in both languages is very similar. Both languages have synthetic and analytic forms. While the synthetic forms can carry all the information in one single lexical unit, their number is significantly small on account of the analytic forms. Word formation of analytic forms in both languages uses main verb and different forms of auxiliary verbs. Main difference in Bulgarian is that only the verb "съм" ("*sam"* be)[1] in its different forms is used to form all the synthetic forms, thus this auxiliary verb carries most of the grammatical information, while in English several other auxiliary verbs combine with each other to form different meanings. Other distinctive feature of the word formation of the verb forms in Bulgarian is that both the main verb and the auxiliary verb can carry grammaticalized information about the grammatical gender of the doer of the action denoted by the verb. These differences contribute to the complicated lexical transfer between Bulgarian and English.

Of course the differences and the similarities of both languages` verb systems are much more numerous and complicated, in our introduction we try to outline the most essential ones, which lead to several difficulties of conducting lexical and semantic transfer regarding machine translation.

3. Differences and similarities of the semantics of the temporal systems of both languages.

As we said, both languages temporal systems share a common feature – they consist of categories within the hyper-category. Both Bulgarian and English have category that expresses a completed action in relation to a referential point – the perfect tenses. Obvious difference is the presence of continuous tenses in English, which can express an action that is uncompleted related to the referential point, as opposed to Bulgarian where such tenses do not exist. Another tangible difference is that the Bulgarian verbs have lexical aspect, which is part of the semantic of the lexical unit and expresses the action as finished or unfinished related to the action`s own completion (Kucarov 2007:551). These two grammatical categories contribute to one of the great difficulties when translating from Bulgarian to English – altering the semantic information of one lexical grammatical category into morphological grammatical. While sometimes changes in meaning are not perceptible, most of the times we have two different meanings: *Чел съм романа/Прочел съм романа- I have read the novel*.

The greater number of possible grammatical categories, therefore possible grammaticalized meaning, in Bulgarian contributes to high levels of ambiguity during translation, due to the fact that in English the possible grammatical categories are less and the grammaticalized information from the source language needs to be reduced or unevenly distributed between different grammatical categories in the target language. Nevertheless, as it has been pointed out before (Lazarov, 2016), the characteristics of grammaticalized information in Bulgarian and English verb forms share numerous similarities. That is why we have similar grammatical meaning in most of the verb forms. We have to point out again that most of the grammaticalized information is lost during the semantic transfer between the grammatical categories of both languages. The grammatical number and person are grammaticalized by every form in Bulgarian and most of the verb forms carry information about the grammatical gender of the doer of the action, whereas in English most of the times we have tenses with only one form. In Table 1 we present as example the formal accordance in meaning between Bulgarian and English tenses and the ratio of the forms.

Bulgarian	**English**	Number of forms
Praesens	**Present simple/Present continuous tense**	
Person, number	3rd person, sg. num./1st person and 3rd person, sg.num	6:2/6:3

[1] The particles from Old Bulgarian language ща (*schta,* will) is also used the word formation of Futurum.

Aorist	**Past simple tense**	
Person, number	-	5:1
Imperfekt	**Past continuous time**	
Person, number	1st and 3rd person, sg. .umn.	5:2
Perfekt	**Present perfect tense /Present perfect continuous tense**	
person, number, gender	3rd person, sg. num.	12 :2
Plusquamperfekt	**Past perfect/Past perfect continuous tense**	
person, number, gender	-	9 :1
Futurum	**Future simple/Future simple continuous tense**	
Person, number	-	6:1
Futurum exactum	**Future perfect/future perfect continuous tense**	
person, number, gender	-	12:1
Futurum praeteriti	**Future simple tense in the past (*going to*)**	
Person, number	1st and 3rd person, sg. .num.	6: 2
Futurum exactum praeteriti	**Future perfect in the past /Future perfect continuous tense in the past**	
person, number, gender	-	12:1

Table 1: Accordance in meaning of tenses between Bulgarian and English.

The huge diversity of verb forms in Bulgarian leads to several problems when translating in English. For the purposes of transfer-based machine translation developing rules for all possible variations, although more reliable, can be time-consuming and hard. On one hand, the much smaller number of forms in English can be a great advantage, because a large number of forms in Bulgarian are transferred into a much smaller in English, thus the possible outcomes in the target language are equal to the number of the transfer rules (Lazarov, 2016). On the other hand, most of the forms, which grammaticalize meaning for evidentiality, voice and mood, have very low frequency and incomprehensible usage. Of course we have to point out that for the purposes of rule-based machine translation this problem can be resolved by providing more precise contextual rules. Our point of view is that these problems can be better studied and resolved by the method of statistical language modeling.

4. Towards the statistical method in machine translation.

As we said before, the rule-based method in machine translation is reliable, as it depends on language models, which are constructed by people – thus the knowledge of language is exterior – it is still the human competence of language. Essential for the rule-based method is the presence of large and accurate grammars and dictionaries, which must take into account all possible language variations. Needless to say, there is no such grammar that can describe human language in such depth and detail in all of its possible manifestations. Therefore we need to gather information about the language not by prescribing it, but by describing its actual usage – we need a grammar that prescribes probable language models, rather than describing theoretical ones.

In the short history of computational linguistics and machine translation we have achieved more than the fathers of this scientific field ever imagined and predicted. Starting from the basic understanding of language as a set of rules, nowadays we have opportunity to discover more and new inner dependencies throughout all natural languages. We are able not only to build grammatical models of languages, but also statistical ones.

The goal of statistical language modeling is to build a statistical language model that can estimate the distribution of natural language as accurate as possible. A statistical language model is a probability distribution P(s) over strings S that attempts to reflect how frequently a string S occurs as a sentence (Song and Croft, 1999: 317). By expressing various language usages and deviations in terms of simple parameters in a statistical model, it can provide an easy way to deal with complex natural

language phenomena. Statistical language modeling (SLM) originated in the late 1980`s for the purposes of speech recognition, but it has also played a vital role in various other natural language applications like machine translation, part-of-speech tagging, intelligent input method, etc. It has passed through two periods of its development – word based SLM (1992) and phrase based SLM (2003). Main principle of statistical language modeling is more data is better data. For the purposes of statistical machine translation (SMT) we need enormous corpora with enough variable data in order to provide enough linguistic material. As we said, both Bulgarian and English are resource rich languages and both can provide sufficient data for research on their own and between them. Figure 1 shows the process of analyzing data in SMT.

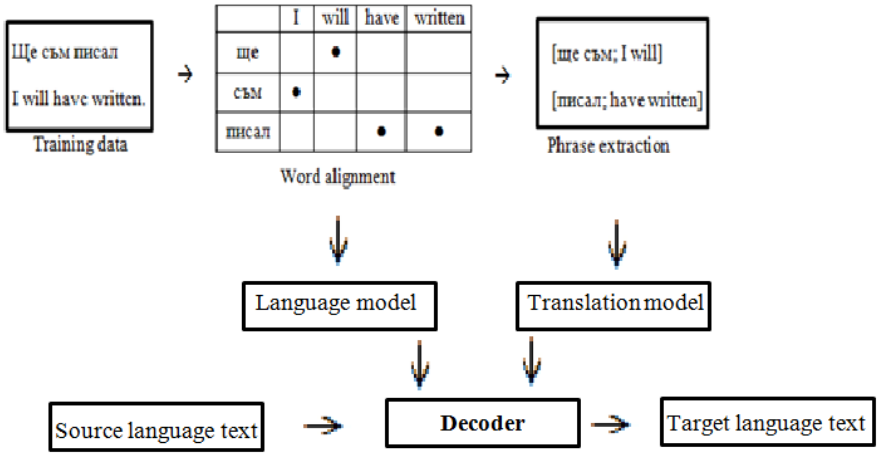

Figure 1: Simplified process of analyzing data from bilingual corpora.

For the purposes of statistical language modeling we need to know whether a given string of words is the right string of words in given language – we need to know what the probability of the string is. That is why we need to decompose this probability to the product of the probabilities of each word appearing in context of other words. Nowadays the n-gram model is the most widely used for language modeling and SMT. In a n-gram model, the probability $P(w_1,w_2,\ldots w_n)$ of observing the sentence $w_1,w_2,\ldots w_n$ is calculated as: $\prod_{i=1}^{m} P(w_i|w_1,w_2,\ldots w_{i-1}) \approx \prod_{i=1}^{m} P(w_i|w_{i-(n-1)}\ldots w_{i-1})$ or it is assumed that the probability of observing the i^{th} word w_i in the context history of the preceding $i-1$ words can be approximated by the probability of observing it in the shortened context history of the preceding $n-1$ words.

Of course statistical language modeling directly depends on quantity and quality of the avilable linguistic resources. Main principle of it is "more data is better data", thus a statistical model of certain language evaluates the probability of certain string of words to appear not by their grammatical correctness, but by the frequency of their usage in the available resources. We need to specify that we can calculate the probability not only of words, but of any linguistic unit – phonemes, morphemes, words, phrases etc. In this paper we introduce the idea that we can build statistical language model of the verb systems of Bulgarian and English adding up two methods of machine translation.

5. A possible approach of applying SMT for the purposes of translating verb forms.

As it has been pointed out, there are formal similarities of the semantics of the temporal systems of Bulgarian and English. Nevertheless, transfer-based rules are not reliable enough to translate the majority of Bulgarian verb forms in English. This is why we propose the hypothesis that a possible collaboration between these two methods (rule-based MT and SMT) can be a solution to this problem.

We already pointed that transfer-based rules can provide information about the exact semantic and lexical transfer between the two languages of interest for us, nevertheless, in the case of translating from Bulgarian to English they cannot be prescribed with 100% certainty due to the huge amount of grammatical information that is lost during the process of translation, thus we need to construct a statistical language model of the transfer-based rules on their own. After that we could generate translation model, which is going to rely on the transfer-based rules. Our view is that using PoS-annotated corpora we can construct language model of the verb systems of both languages, which

language models will be able to present statistical information about the usage of different morphological categories and the frequency of some of the verb forms that have uncertain usage and vague meaning. After constructing these single language statistical models, we will be able to construct the statistical translation model of the transfer-based rules. By comparing the data from extracted verb forms from parallel corpora, we can see how frequently the data we have supports the usage of given transfer-based rules. In that way we will be able even to derive new rules, if the present ones do not get approval by the available linguistic data. The final stage of constructing the translation model is going to include verification of the gathered data by attempting to translate different types of other textual resources.

As we know the phrase based translation gives us better results, so as the verbal phrase in sentences constructs a whole syntactical unit, we can try to analyze the whole VP and build a language model based on its behavior. As we said before the asymmetric grammatical categories in English and Bulgarian contribute to the fact that large number of forms in Bulgarian have to be translated with much smaller number of forms in English. We propose that creating a language model of the verbal systems of both languages for the purposes of machine translation can be achieved by implementing the transfer-based rules with statistical language models. Our hypothesis is that a statistical language model of the verbal systems of both languages can complete the language model that the rule-based method composes. As we said, the rule-based method gives us strict information about the semantic and lexical transfer from one language to another, but in our case we have more coinciding verb forms in English for the Bulgarian forms. Taking into account what is the probability of certain verb forms to occur in English when we have a given forms in Bulgarian, we can prescribe our transfer-based rules with this certain probability. In this way we can have information based on actual data of language usage combined with exterior knowledge of language. Combining these two methods, we can relate verb forms with certain probability between languages. Also in cases where two or more verbal forms in English correspond to one in Bulgarian we will have statistical data of the probability of each corresponding form to occur in our target language and the context in which it can occur. This can help us theoretically establish any correlation between the lexical aspect of the verb in Bulgarian and the category of aspect in English. Another aspect in which implementing statistical and rule based machine translation can help us is to establish, based on various data, what is the statistical probability of certain verbal form to occur – as we know the verbal forms for evidentiality, mood and voice in Bulgarian tend to peter out at the expense of other more frequent forms which carry less grammatical information, but are more recognizable for the users of the language. The lost grammatical information is retrieved within the frames of the sentence. Thus if we get low probability for given verb form, we need better transfer-based rules.

By applying statistical language modeling to the rule based method, we can extract information about given language on its own. We can gain statistical information about the frequency of a certain verb forms in different kinds of texts, thus prescribing a probability of some verb forms to be in this kind of text. Based on the size and quality of the corpora we have, we can make conclusions about what type and what size of grammatical information is lost during the translation process, so in future we can try to figure out ways to prevent that by providing more contextual rules. This way we can also gather information about the cases in which we lose grammaticalized information because of dissimilarities in the working languages. A simplified chart of our linguistic model is presented in Figure 2.

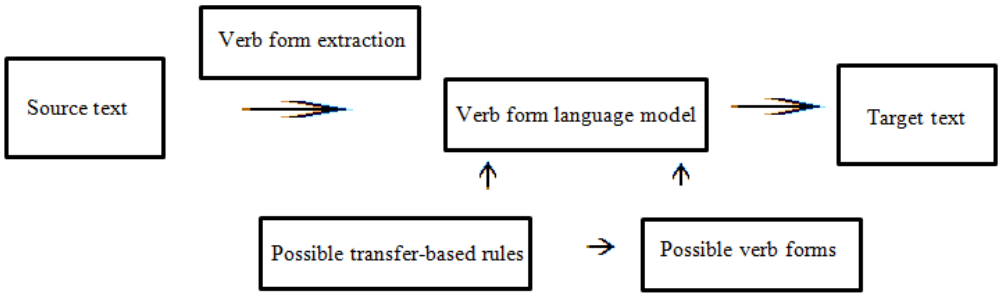

Figure 2: Chart of the stages of analyzing the verb forms in two languages.

6. Conclusion and future research

Of course we have to point out that our idea, although restricted to analyzing only the verb forms in Bulgarian and English, carries the possibility to give us new look at the linguistic data that we have of both languages. It is dependable on the quality and quantity of the corpora we have, it will also depend on the transfer rules we already have. The perfect corpora of Bulgarian and English for the purposes of our research must have aligned sentences with several types of annotations – morphological, syntactical, semantical and also information about the possible transfer rules and their variations. The available Bulgarian National Corpus, the Bulgarian PoS annotated corpus and the British National Corpus are suitable for constructing the preliminary single language models, in order to gather data about the frequency of usage of the verb forms. For the purposes of constructing the translation model we will need parallel corpora with PoS annotation such as the Bul-X-Cor and also other parallel language corpora will be suitable after careful PoS-annotation. Our future research will include, first of all, gathering and analyzing the available corpora. Extracting all the verb phrases and the context in which they appear. After we analyze this information, we can continue with constructing our verb language model, which must also include information about all possible derivations from the available data we have. The final stage of our work will include comparison of the two fundamental methods we use – transfer-based and statistical, in order to find out what kind and what number of mistakes each of them makes and how they piece out. In that way combining the two main methods of machine translation – rule based and statistical, we will be able to study English and Bulgarian verb systems on their own and also to find the deep inner dependencies between both languages that are in the middle of our linguistic competence and performance.

References

Gerdzhikov, G. (2000). Kategoriyata vreme kato hiperkategoriya. *Bulgarian language and literature – educational journal*, Ministry of Education and Science. http://liternet.bg/publish/ggerdzhikov/hyper.htm

Hutchins, W. J. (1986). *Machine Translation: Past, Present, Future*. Ellis Horwood, Chichester, UK. Halstead Press, New York.

Kucarov, I. (2007). *Teoretitshna gramatika na balgarskiya ezik. Morfologiya.* University Press "Paisii Hilendarski".

Lazarov, T. (2016). Osobenosti na glagolnite sistemi i natshinite za izrazyavane na vremeto v balgarski i anglijski. Semantitshen transfer pri prevod ot balgarski na anglijski. *Littera et Lingua – electronic journal*, Sofia University. http://slav.uni-sofia.bg/naum/en/lilijournal/2015/12/1-2/tlazarov

Nitsolova, R. (2008). *Balgarska gramatika. Morfologiya*. University Press "St. Kiment Ohridski".

Song, F. and W. B. Croft (1999). A General Language Model for Information Retrieval. In *Proceedings of the Eight International Conference on Information and Knowledge Management.* ACM, New York, USA.