EFFICIENT AND HIGH-QUALITY SOLUTIONS FOR MONO- AND MULTILINGUAL ISSUES BASED ON METHODS OF STATISTICAL MACHINE TRANSLATION

THESES OF THE PH.D. DISSERTATION

László János Laki

Scientific advisor:

Gábor Prószéky, D.Sc.

doctor of the Hungarian Academy of Sciences

Pázmány Péter Catholic University

Faculty Of Information Technology And Bionics

Multidisciplinary Technical Sciences

Doctoral School

Budapest, 2015.

1. Introduction and research aims

One of the most important tasks of human language technology is to bridge the barriers of language diversity, namely to qualify computers for translation between different languages. Over the past few years, the explosive growth of information technology has enabled computational linguists to provide a solution for this problem. Currently, the most widespread method is to train statistical machine translation (SMT) systems. An SMT system is a purely language-independent tool, which can be trained by unsupervised machine learning algorithms, and the quality of the translation is acceptable if using an appropriate amount of training data. The disadvantage of the method is that the purely statistical method is not sufficient to make high quality translations in the case of grammatically distant languages, or between morphologically rich ones. The main problems are derived from the differences of word order and the average number of words in the sentences. The translation to an agglutinative language is difficult due to the frequent need of long distance word reordering at decoding time and the low representation of rare word forms in the training corpora. **The first part of my work deals with solving these difficulties by the hybridization of the translation system. My goal was to develop an architecture that can reduce the negative effects caused by the data sparseness problem, and is able to generate grammatically correct surface word forms.**

Understanding and analyzing of written texts is essential for further text processing. The first step of the processing chain is that of the complete morphological disambiguation, which contains lemmatization – finding the dictionary form of the words – and part-of-speech (PoS) tagging. This task is not trivial, because there are a number of word forms that belong to several morphosyntactic groups. In these cases we can determine the correct group by considering the context of the given word and the position of the word in the sentence. Many applications exist to solve this problem, but only few of them can determine the lemma and PoS tags of the words simultaneously, and even fewer of them is able to work in a language-independent manner. Since an SMT system is supposed to transform two languages into each other, this architecture may be applicable for complete morphological disambiguation. In this case the task is the translation between plain text and a lemmatized—PoS-tagged one. **In the second part of my work my goal was to develop and present a language-independent SMT-based complete morphological disambiguation tool that reaches or exceeds the results of existing language dependent- and independent systems.**

2. Methods and tools

In this work, applications of the methods of *statistical machine translation* (SMT) for different purposes were investigated. I also dealt with further developing existing systems and improving the quality of their results. The basic idea of statistical machine translation is that the system learns the models required for translation from parallel bilingual corpora using unsupervised learning algorithms. A *parallel bilingual corpus* consists of pairs of sentences in the source and target languages. The simple and fast implementability of the algorithm, and its ability for language-independent behaviour have made this method one of the most cited SMT architectures. In this work the most popular implementation of the SMT method was used, namely the MOSES [1] toolkit, which collects the existing tools and implemented algorithms used for machine translation. The SMT architecture produces good results in the case of language pairs with similar syntactic structure and word order. In contrast, phrase-based models can only hardly handle considerable grammatical differences. In this work a hybrid machine translation system is presented that employs syntactically- and morphologically-motivated pre- and postprocessing steps into the SMT workflow. The effects of the development is presented using the results of English to Hungarian translation. The statistical decoder in the improved architecture was replaced by a morphological generator; i.e. the HUMOR [2] generator.

The evaluation of the system is an important task. The basic automatical evaluation method for SMT systems is based on the comparison of the translated and the reference sentences along different features. Nowadays, the most popular evaluation method is BLEU (BiLingual Evaluation Understudy) [3], which offers the solution for the problem of different word order. The essence of the method is that it searches for the phrases of the translation in the reference sentence. The more similarity between the two sentences the more BLEU points it receives. Despite the advantages of the BLEU metrics, several publications emphasize that in many cases BLEU scoring does not correlate with human evaluation. Therefore, the most important architectures presented in this work were examined by human evaluators.

Statistical machine translation is applicable not only for translation between natural languages, but also for transformation between arbitrary texts. My dissertation showed that the task of complete morphological disambiguation could be defined as a translation problem. A suffix-tree based guesser ([4]–[6]) and a morphological annotator tool (HUMOR) were also integrated into the SMT-based system.

3. New scientific results

The results presented in my dissertation can be classified into two thesis groups. The first one presented the improvements of the purely statistical machine translation between grammatically distant languages using hybridization. In the second one, a statistical machine translation based complete morphological disambiguation tool was demonstrated.

THESIS GROUP I.

In thesis group I. was looking for methods to solve the difficulties appearing during the translation of agglutinative languages. The main problems arising during this task are the data sparseness problem and the generation of the surface word form using statistical methods. Further challenge is the translation between grammatically distant languages, since there are considerable differences in word order and in the average number of words in the sentences. In my work, the pure word-based statistical machine translation system was supplemented with algorithms to handle the grammatical differences between source and target languages. The integration of my algorithms into the translation process resulted in better translation quality.

Thesis 1: ***I achieved the hybridization of the purely statistical machine translation system by creating language pair dependent reordering rules defined by the grammatical specialities causing word order differences. Due to the reordering of the source text, a significant improvement in the quality of the translation was measured.***

Published in: [Laki_1], [Laki_4], [Laki_8]

In Thesis 1 I have shown that statistical machine translation systems by themselves are not sufficient to solve the translation between grammatically distant language pairs, since the generally used decoder implementations are only able to do local reorderings. Therefore, I built a hybrid machine translation system that applies syntactically motivated reordering rules on the English source sentences as a preprocessing step. The structures which are to be reordered were find based on constituent and dependency analysis of the sentence. The reordering rules used in the English-Hungarian translation process improved the quality of the system compared to that of the baseline.

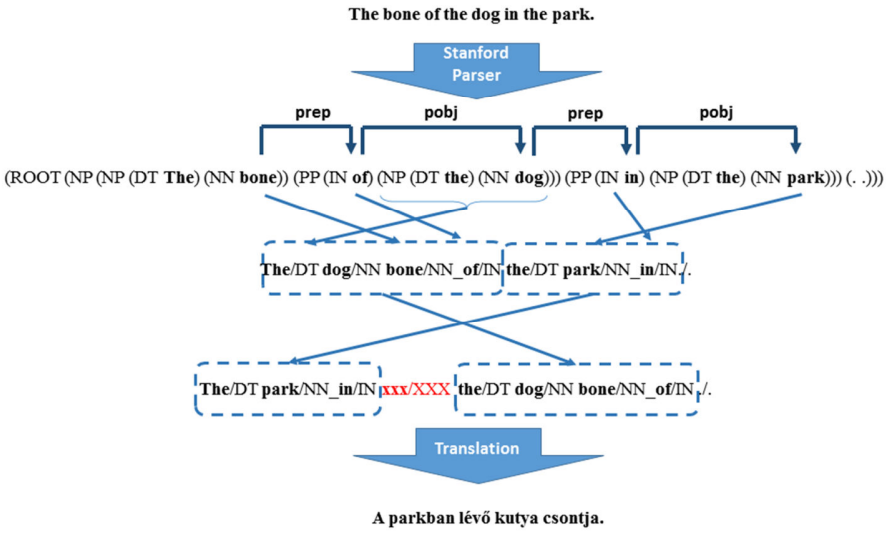


Figure 1: The example shows the reordering of prepositions in the English sentence. When a preposition is part of a possessive structure, it must be translated as a "lévő" structure in the Hungarian sentence. Based on this, the propositional structure (PP) was placed before the possessive noun phrase (NP). An "xxx" string was inserted between these structures, which represents the Hungarian word "lévő". The possessive NP structure was reordered based on a previous rule.

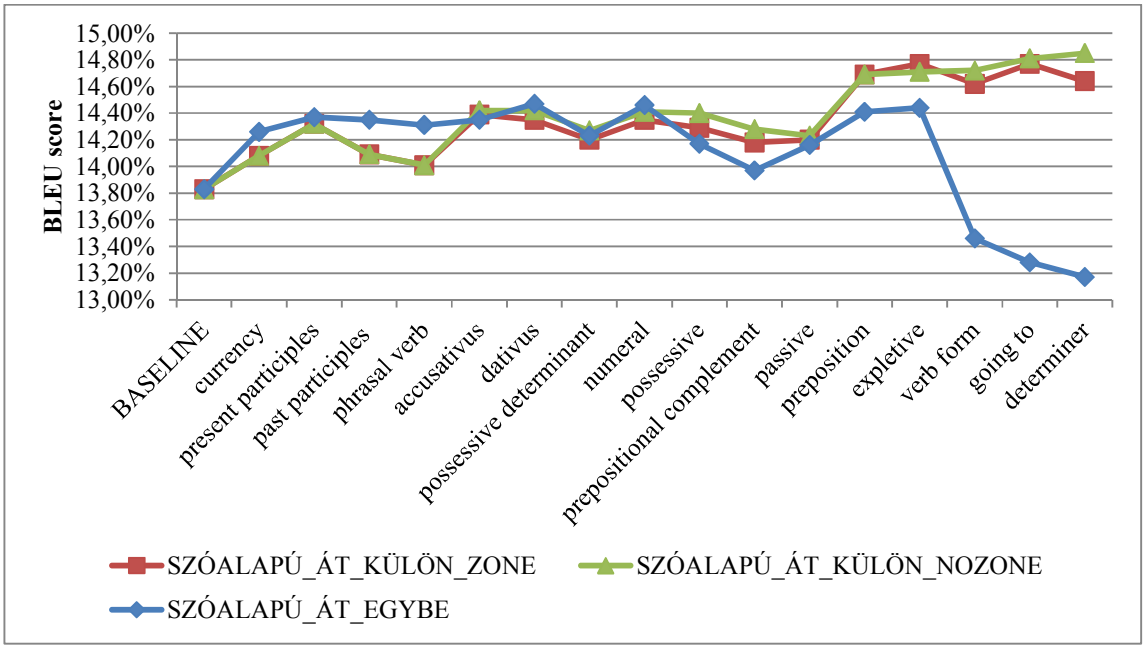


Figure 2: It describes the addition order of the reordering rules and the BLEU score of the systems. The points of the diagram show the systems, which uses all the previously presented rules. The decoder of the MOSES system allows us to define word sequences, which have to be translated together (*SZÓALAPÚ_ÁT_KÜLÖN_ZONE*). The opposite of this system is named *SZÓALAPÚ_ÁT_KÜLÖN_NOZONE* where free translation order is allowed. The system *SZÓALAPÚ_ÁT_EGYBE* joined the English "affixes" (eg. propositions or possessive determinants) to the lemmata.

The best BLEU score using reordering rules was 14.85%, which means 7.38% relative improvement compared to the *SZÓALAPÚ* baseline system (13.83% BLEU score). In addition, many phenomena could be correctly translated this way, not handled in a conventional SMT system. Of course, handwritten rules do not cover all of the phenomena causing word order differences.

Thesis 2: ***I established a translation chain in which the morpheme-based statistical machine translation system was complemented with a morphological generator. In order to solve the problem of homonyms, which occurs frequently in Hungarian, a lemma and morphosyntactic tag based representation was used in the translation workflow instead of the surface form of the words.***

Published in: [Laki_1], [Laki_4], [Laki_8]

Translating into agglutinative languages is a difficult task for the statistical based decoder due to the data sparseness problem. This is the main reason why the quality of the SMT based translation into agglutinative languages stay far below the quality of translation into other languages.

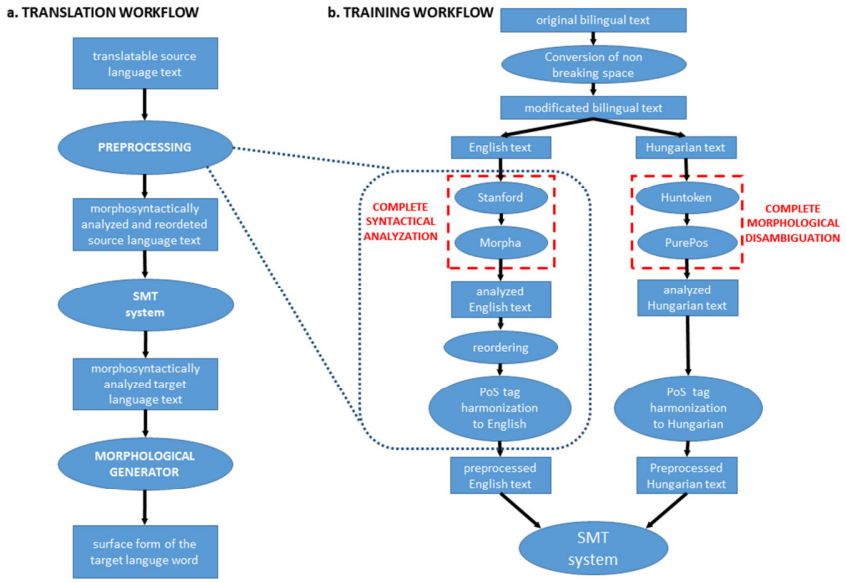


Figure 3: Presentation of the translation process created by the morphological generator and the training process complemented with preprocessing steps

Generating the surface word forms of morphologically rich languages is a complicated task for the decoder, because it is not able to generate a surface word form not represented in the training corpus. I have created a unique hybrid machine translation architecture, in which the SMT-based system is trained on morphologically analyzed source and target texts. The surface word form is created by a morphological generator tool (Figure 2). In contrast to the statistical decoder, the morphological generator is able to create affixed surface word forms with high accuracy, which are not included in the training set. To decrease the effect of data sparseness and homonymy, I used morphosyntactic tags instead of the surface form of morphemes. This way, the number of the possible word forms were reduced greatly. The results of the architectures using the morphological generator constructed in this thesis, are more understandable for human evaluation and they were able to generate more fluent translations compared to the word-based statistical decoder.

Thesis 3: ***I developed a word number harmonisation algorithm working on morphologically parsed source and target texts in which the different morphological behaviors of the two languages were handled by approximating the number of the morphemes and matching them during the translation process. I showed that an improvement can be reached in the translation quality using the word number harmonisation in the case of morphologically complex languages.***

Published in: [Laki_1], [Laki_4], [Laki_8]

Three SMT system architectures are presented in this thesis, which are able to manage the difference in the number of words in the source and target language sentences. First, a word-based SMT system is described that works on morphologically analyzed texts, and allows to English to behave like an agglutinative language. Second, a morpheme-based SMT system is demonstrated that translates between texts splitted to morphemes. The third system is a factored SMT, which combines the advantages of the previous two systems. This method translates from lemma to lemma and from morphemes to morphemes simultaneously. The system is unique, because instead of full word forms, the output of the system is a record composed of the lemma and the related PoS tags of the word from which the surface word form is created by a morphological generator. (This morphological generator was described in Thesis 2.)

I showed a solution for the problem arising from different word number in Hungarian and English sentences with the help of morpheme-based translation (word-, morpheme- and factored translation models). Automatic evaluations according to the BLEU metrics were performed at several phases of my work, and in some cases human evaluations were also carried out. The measurements confirmed that lower BLEU scores do not necessarily mean lower translation quality. Thus, it was proved that the systems built using word number harmonisation reached better translation quality according to human evaluators than the purely statistical ones (Table 1).

System name	Human evaluation	w-BLEU	mm-BLEU
reference	88,33%		
MetaMorpho	76,30%	6,86%	50,97%
Google Translate	72,80%	15,68%	55,86%
Bing Translator	61,66%	12,18%	53,05%
MORFÉMAALAPÚ_ÁT_T6	**55,60%**	12,22%	**64,94%**
FAKTORÁLAPÚ_ÁT_T6_FIX	55,42%	10,88%	60,83%
MORFÉMAALAPÚ_T6	54,28%	12,19%	63,87%
FAKTORÁLAPÚ_T6_FIX	52,03%	9,91%	57,09%
SZÓALAPÚ_T6	51,33%	13,83%	59,32%
SZÓALAPÚ_ÁT_T6	50,89%	**14,83%**	58,06%
SZÓALAPÚELEMZETT_ÁT_T6	37,57%	13,05%	57,21%

Table 1: Human and machine (BLEU) evaluation of translation systems including morphological amendment and their comparison to generally used translation systems. According to the table, human and machine evaluation are not in accordance with each other. Despite the lower BLEU score, my system architecture in many cases (*MORFÉMAALAPÚ_ÁT_T6, FAKTORÁLAPÚ_ÁT_T6_FIX, MORFÉMAALAPÚ_T6, FAKTORÁLAPÚ_T6_FXI)* achieve better performance for the human reader than the baseline (*SZÓALAPÚ_T6*).

Thesis 4: ***I showed that the quality of the translation improves if the training set is augmented with a bilingual vocabulary containing accurate translation of short phrases (e.g. lexical items, example dictionary). Taking these phrases into account with adequate weights, balance the statistics calculated from the training set containing longer sequences, and also makes the translation model more robust.***

Published in: [Laki_11], [Laki_12]

The word alignment often does not find the corresponding text fragments, because they are far from each other due to the grammatical structure of the languages or if they are very different from each other. There are also problems with too long sentences, because in many cases the words in

the target clause are joined to one of the source words, or frequent words which are represented several times in the sentence couldn't find their correct pairs. To solve these problems, I have complemented the training set with a bilingual corpus containing accurate translations of short phrases. The created dictionary has been added to the training set several times, because the precise phrases has to be more significant in the translation model. Therefore, the relevance of the original corpora and the weighting of multiword phrases is decreased in the translation model. Furthermore, the quality of the language model is also reduced.

Based on the experiments, the best result could be achieved when the dictionary is added to the learning set only once. In this case the *ALAPRENDSZER* shows a 0.33% BLEU score improvement reaching 11.18% instead of 10.85%, as shown in Table 2.

System name	BLEU score
ALAPRENDSZER system	10,85%
ALAP+1XSZÓTÁR system	**11,18%**
ALAP+2XSZÓTÁR system	11,01%
ALAP+3XSZÓTÁR system	10,88%
ALAP+4XSZÓTÁR system	10,87%
ALAP+5XSZÓTÁR system	10,87%

Table 2: The results of the systems complemented with a dictionary

THESIS GROUP II.

In the second part of my dissertation, I have presented a new approach for complete morphological disambiguation based on SMT. The unique property of this system is that it makes lemmatization and PoS tagging simultaneously. In addition, thanks to the language independent modules, the system could be applicable to any kind of languages or PoS tag sets. The evaluation has proved that the performance of my system is comparable to and in many cases it outperforms the results of existing language-independent systems, and even approximates the performance of some language-dependent ones. Due to the effective handling of unknown words (OOV), I have integrated a suffix-tree based morphological guesser into the workflow. Finally, the performance of the system was increased due to the integration of the guesser and the morphological analyzer combination. I have investigated the efficiency of the method in several PoS tag sets and languages (English, Portuguese, Bulgarian, Hungarian, Croatian and Serbian).

Thesis 5: ***I developed a complete morphological disambiguation tool, which is able to make lemmatization as well, based on SMT methods. I showed that decreasing the number of PoS tags in the target side tagset considerably improves the quality of the system.***

Published in: [Laki_2], [Laki_3], [Laki_5], [Laki_6], [Laki_7], [Laki_9], [Laki_10], [Laki_12]

Since the statistical translation system transforms a language into another one, therefore it could be used for translating between plain texts and annotated ones. In my dissertation I uniquely exploited this property; a complete SMT-based morphological disambiguation system was created that performs lemmatization and Part-of-Speech tagging simultaneously. Accurate complete morphological disambiguation is essential for the processing of agglutinative languages.

For the statistical system, the relevance of the contextual information is weakened due to the sophisticated specification of tags, thus morphosyntactic disambiguation could be resolved harder. The solution for this problem is decreasing the number of items in the target dictionary by the generalization of the tagset. In my work, I have implemented and compared several methods to reduce the target tagset. The first technique to reduce the complexity of the system was simplifying the information stored in PoS tags.

It means the abandonment of the lemmatization and the less important PoS subclasses from the original PoS tags. Despite the large amount of information loss, the quality of the annotation system is getting better in the case of the (*ALAP_SZIMB_SZÁM_CSAKPOS, ALAP_SZIMB_SZÁM_FOPOS*) systems. The second solution is able to save not only the stored information, but is also able to reduce the complexity of the disambiguation system. The solution was to store the target stems in a more compact form. Similarly to the results of Orosz and Novák [6], I also represented the lemmas as a record, which gives the required transformation needed to be done on a given word to get its lemma. For example, in the case of the record <*delete,paste*> , where the *delete* gives the number of the characters that have to be deleted from the end of string, and the *paste* represents the string that has to be joined to the end of the "mutilated word" to get its lemma. The system constructed this way (*TÖRÖLCSATOL_SZIMB_SZÁM*) is more effective in lemmatization and PoS tagging than the baseline one (Table 3). According to the results achieved, it was proved that the performance of the disambiguation system would improve, if the complexity of the target PoS tag set were reduced.

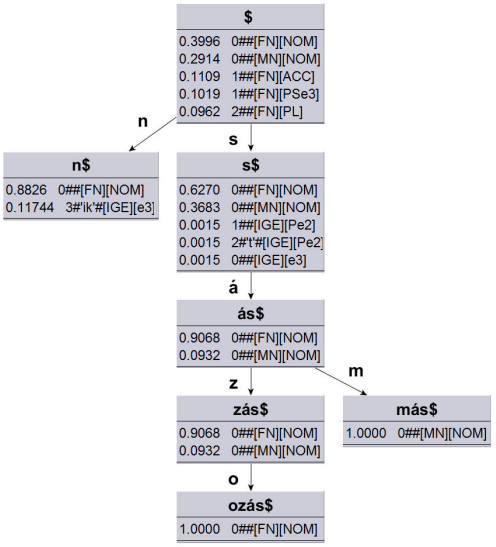
	Word level			Sentence level	
	PoS	Lemma	Complete	PoS	Complete
ALAP	91,281%	94,303%	91,257%	35,371%	35,294%
ALAP_SZIMB_SZÁM_CSAKPOS	91,534%	-	91,534%	37,071%	37,071%
ALAP_SZIMB_SZÁM_FOPOS	95,471%	-	95,471%	53,898%	53,898%
TÖRÖLCSATOL_SZIMB_SZÁM	91,496%	94,330%	91,447%	36,977%	36,684%

Table 3: Results of systems constructed with the reduce of target tag set

Thesis 6: ***To handle OOV words in the training set, I integrated a suffix-trie based morphological guesser into the SMT based disambiguation system. This module helped my system to outperform the available language independent tools.***

Published in: [Laki_2], [Laki_3], [Laki_6]

The main fault of disambiguation systems lies in the analyzation of out-of-vocabulary (OOV) words. It is especially true in the case of agglutinating languages, where one stem could have up to hundreds of different inflected surface forms, but not all of them are included in the training set. Thus, the disambiguation system has no prior knowledge of these words. Based on the experiences presented in this thesis, the OOV words could be modeled using rare words in the training set. The first method supports the SMT-based disambiguation system with classification of OOV words into bigger word classes. The method assumed that OOV words behave similarly to those having similar positions in the sentences in the training set. The PoS tag of an OOV word can be inferred from the position and the inflections of the word in the sentence or rather from surrounding words and their PoS tags. Therefore, OOV words were replaced by an *unk_abcd* format (hereafter *unk-suffix*), where *abcd* means the last four characters of the word. The potential information from rare words was used during the training of the system, since the behaviour of unknown words was modeled using rare words. In practice, this means that rare words in the training set were replaced by the unk-suffix string. However, the disadvantage of this system is that only suffixes of fix length are taken into account during disambiguation.



$$P(0\#\#[FN][NOM]|facebookozás) = \theta_0 P(0\#\#[FN][NOM]) \times \theta_1 P(0\#\#[FN][NOM]|\text{”s”}) \times \theta_2 P(0\#\#[FN][NOM]|\text{”ás”}) \times \theta_3 P(0\#\#[FN][NOM]|\text{”zás”}) \times \theta_4 P(0\#\#[FN][NOM]|\text{"ozás"})$$

Figure 4: The presentation of searching in the suffix-trie based guesser

To solve the deficiency of fix suffix length, a suffix-guesser was integrated to the disambiguation chain. There is an opportunity in the Moses framework to define translation options that are taken into account by the decoder during translation. This way a pre-translation can be defined to OOV words as a proposal of the guesser. The used guesser builds a suffix-trie from the words of the training set, where the information, which includes the probability of different PoS tags in the case of the given suffix, is stored in the nodes. These probabilities were trained based on rare words in the training set. As a result, the system complemented with the guesser improved the accuracy of disambiguation of OOV words considerably (Table 4).

	Word level		Sentence level		
	PoS	Lemma	PoS	Lemma	PoS
TÖRÖLCSATOL_SZIMB_SZÁM	91,496%	94,330%	91,447%	36,977%	36,684%
TÖRÖLCSATOL_SZIMB_SZÁM_ UNKSZUFFIX	96,025%	97,828%	95,383%	58,752%	54,284%
TÖRÖLCSATOL_SZIMB_SZÁM_ GUESSER	96,511%	98,595%	96,177%	62,465%	59,692%

Table 4: Results of the systems constructed with technics handling OOV words

Thesis 7: ***I presented the language independent behaviour of SMT based disambiguation systems. Therefore, my tool was trained with seven different languages and morphosyntactic code sets. The results were comparable with the quality of the systems available for the given languages.***

Published in: [Laki_3], [Laki_6]

Results of the language-independent complete morphological disambiguation tool created in this work was compared with the performance of available ones for other languages and code sets. The experiments revealed that my system is comparable to others - sometimes also language dependent ones -, and in many cases even exceeds them (Table 5).

Language	System name	Word level accuracy		
		PoS	lemma	complete
Hungarian (HUMOR)	PurePos	96,50%	96,27%	94,53%
	HuLaPos2	96,70%	98,23%	97,62%
	PurePos+MA	98,96%	99,53%	98,77%
Croatian	**HuLaPos2**	**93,25%**	96,21%	90,77%
	HunPos+CST	87,11%	**97,78%**	-
Serbian	**HuLaPos2**	**92,28%**	92,72%	86,51%
	HunPos+CST	85,00%	**95,95%**	-
Bulgarian	TnT [4]	92,53%	-	-
	machine learning	95,72%	-	-
	machine learning +morf. lexicon	97,83%	-	-
	HuLaPos2	97,86%	-	-
	machine learning +morf.lexicon+rule	97,98%	-	-
Portuguese	**HuLaPos2**	**93,20%**	-	-
	HMM based PoS tagger	92,00%	-	-
English	TnT [4]	96,46%	-	-
	phrase based translator[12]	96,97%	-	-
	HuLaPos2	97,08%	-	-
	Stanford tagger 2.0 [7]	97,32%	-	-
	SCCN [8]	97,50%	-	-

Table 5: Comparison of the results of different complete morphological disambiguation tools in several languages

Thesis 8: ***I showed that the integration of a morphological analyzer into the SMT workflow of my system results better accuracy for text annotation.***

Published in: [Laki_3]

This thesis proves that the language independent morphological disambiguation tool completed with language-specific morphological analyzer results in further quality improvements. Using this method, one of the most accurate systems for Hungarian has been created, which performs lemmatization with an accuracy of 99.12%, in addition, 84.82% of the OOV words were tagged correctly, and the accuracy of the complete morphological disambiguation was 96.50% (Table 6).

	Word level		Sentence level		
	PoS	Lemma	PoS	Lemma	PoS
TÖRÖLCSATOL_SZIMB_SZÁM_UNKSZUFFIX	96,025%	97,828%	95,383%	58,752%	54,284%
TÖRÖLCSATOL_SZIMB_SZÁM_GUESSER	96,511%	98,595%	96,177%	62,465%	59,692%
TÖRÖLCSATOL_SZIMB_SZÁM_MORFLEXIKON	96,624%	**99,119%**	**96,498%**	63,236%	62,250%
PUREPOS2	96,350%	97,505%	95,101%	60,817%	51,294%
HUNPOS+CST_SZÓTÖVESÍTŐ	96,340%	96,512%	94,276%	61,279%	47,288%
MORFETTE	**96,751%**	96,048%	93,776%	64,591%	44,160%
NLTK_MAXENT+ CST_SZÓTÖVESÍTŐ	94,949%	95,439%	92,927%	51,402%	40,169%

Table 6: Comparison of the results of different systems available on Hungarian

4. Application of the results

The research described in my dissertation focused on problems that first contribute to the improvement of translation between languages, and second to the refining of complete morphological disambiguation. Results related to hybrid machine translation can be successfully integrated into architectures of arbitrary SMT systems. The results confirmed that using morphological information in the translation workflow increases the quality of the translation.

The complete morphological disambiguation tool presented in the second thesis group is able to operate in language dependent and independent manners. The described method is suitable for the integration into syntactical parsing chains. Furthermore, according to Orosz et al. [Orosz_1, Orosz_2], the SMT-based disambiguation tool is appropriate for clarifying the results of disambiguation tools operating along different principles, with the combination of such tools.

5. Acknowledgment

First of all, I would like to thank to my consultant, Professor Gábor Prószéky, D.Sc., for his guidance and support during the last few years. I am thankful for the professional governance and also raising my awareness to lectures, conferences and publication opportunities related to my research. Thanks him for the friendly consideration, which was shown towards me all along. The good point and the value of my work was noticed by him. This work could not be established without him.

I am thankful for the former and current leaders of Multidisciplinary Technical Sciences Doctoral School of Pázmány Péter Catholic University, especially deans Tamás Roska, D.Sc., Judit Nyékyné Gaizler, D.Sc. and Péter Szolgay, D.Sc., that the opportunity to my Ph.D. work in the Faculty was given.

I would like to thank to doctors and professors of the Catholic University of Leuven, mainly to Vincent Vandeghinste, Frank Van Eynde and Ineke Schuurmann, for the inspiration and the introduction into the world of statistical machine translation.

Thanks to my closest colleagues for the professional and friendly support that have been perceptible during my doctoral years. I would like to thank primarily to my co-authors, Borbála Siklósi, György Orosz and Attila Novák, who have helped me during my research and the preparation of my publications. Special thanks to Nóra Wenszky, Ph.D., for the English and Hungarian proofreading. Additional thanks to the members of Natural Language Processing Group of PPKE ITK, István Endrédy, Balázs Indig, Márton Miháltz, Ph.D., Bálint Sass, Ph.D. and Győző Yang Zijian, for the brainstormings and cheerful atmosphere.

Thanks for the friendly discussions and encouragement to my previous and present colleagues at the Doctoral School fellows, particularly András Laki, András Bojárszky, Gergely Feldhoffer, Ph.D., Tamás Fülöp, László Füredi, András Gelencsér, Domonkos Gergelyi, András Horváth, Ph.D., András Kiss, Ph.D., Miklós Koller, Ph.D., Dániel Kovács, Csaba Nemes, Ph.D., Tamás Pilissy, Mihály Radványi, Ádám Rák, Ph.D., Attila Stubendek, Antal Tátrai, Ph.D., Róbert Tibold, Ph.D., Dávid Tisza, Gábor Tornai, Ph.D., Kálmán Tornai, Ph.D., Emília Tóth and Tamás Zsedrovits, Ph.D.

I am really thankful to the administrative and the financial staff of the Faculty who dealt with my problems helpfully and flexibly.

Last but not least, I would like to thank to my whole family, that they always helped, encouraged and supported me in every possible way during my research.

6. List of publications

Journal publication of the author:

[Laki_1] **Laki, László János**, Attila Novák, and Borbála Siklósi. 2013. "Syntax Based Reordering in Phrase Based English-Hungarian Statistical Machine Translation." *International Journal of Computational Linguistics and Applications* 4 (2): 63–78.

Book chapter:

[Laki_2] **Laki, László János**, György Orosz, and Attila Novák. 2013. "HuLaPos 2.0 – Decoding Morphology." In: *Advances in Artificial Intelligence and Its Applications*, edited by Félix Castro, Alexander Gelbukh, and Miguel González. Lecture Notes in Computer Science Vol. 8265, 294–305. Springer: Berlin-Heidelberg.

International conferences of the author:

[Laki_3] **Laki, László János**, and György Orosz. 2014. "An Efficient Language Independent Toolkit for Complete Morphological Disambiguation." In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 26–31. Reykjavik, Iceland: European Language Resources Association (ELRA).

[Laki_4] **Laki, László János**, Attila Novak, and Borbála Siklósi. 2013. "English to Hungarian Morpheme-Based Statistical Machine Translation System with Reordering Rules." In: *Proceedings of the Second Workshop on Hybrid Approaches to Translation*, 42–50. Sofia, Bulgaria: Association for Computational Linguistics.

[Laki_5] **Laki, László**. 2012. "Investigating the Possibilities of Using SMT for Text Annotation." In: *1st Symposium on Languages, Applications and Technologies*, 21:267–283. OpenAccess Series in Informatics (OASIcs). Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

Hungarian conferences of the author:

[Laki_6] **Laki, László János**, and György Orosz. 2014. "HuLaPos2 - Fordítsunk morfológiát." In: *X. Magyar Számítógépes Nyelvészeti Konferencia*, 41–49. Szeged: Szegedi Egyetem.

[Laki_7] **Laki, László János**, and György Orosz. 2013. "Morfológiai egyértelműsítés nyelvfüggetlen annotáló módszerek kombinálásával." In: *IX. Magyar Számítógépes Nyelvészeti Konferencia*, 331–337. Szeged: Szegedi Egyetem.

[Laki_8] **Laki, László János**, Attila Novák, and Borbála Siklósi. 2013b. "Hunglish mondattan – átrendezésalapú angol-magyar statisztikai gépifordító-rendszer." In: *IX. Magyar Számítógépes Nyelvészeti Konferencia*, 71–82. Szeged: Szegedi Egyetem.

[Laki_9] **Laki, László János**. 2012. "SMT módszereken alapuló szófaji egyértelműsítő és szótövesítő rendszer." In: *VI. Alkalmazott Nyelvészeti Doktorandusz Konferencia*, 121–133. Budapest: MTA Nyelvtudományi Intézet.

[Laki_10] **Laki, László János**. 2011a. “Statisztikai gépi fordítási módszereken alapuló egynyelvű szövegelemző rendszer és szótövesítő.” In: *VIII. Magyar Számítógépes Nyelvészeti Konferencia*, 12–23. Szeged: Szegedi Egyetem.

[Laki_11] **Laki, László János**. 2011b. “Angol-magyar statisztikai gépi fordító rendszer minőségének javítása.” In: *V. Alkalmazott Nyelvészeti Doktorandusz Konferencia*, 77–86. Budapest: MTA Nyelvtudományi Intézet.

[Laki_12] **Laki, László János**, and Gábor Prószéky. 2010. “Statisztikai és hibrid módszerek párhuzamos korpuszok feldolgozására.” In: *VII. Magyar Számítógépes Nyelvészeti Konferencia*, 69–79. Szeged: Szegedi Egyetem.

Other publications of the author:

[Laki_13] **Laki, László János**, and György Orosz. 2011. “VII. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, 2010. December 2–3.” *Magyar Terminológia* 4: 119–123.

[Orosz_1] Orosz, György, **László János Laki**, Attila Novák, and Borbála Siklósi. 2013. “Combining Language Independent Part-of-Speech Tagging Tools.” In: *2nd Symposium on Languages, Applications and Technologies*, edited by José Paulo Leal, Ricardo Rocha, and Alberto Simões, 29:249–257. OpenAccess Series in Informatics (OASIcs). Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[Orosz_2] Orosz, György, **László János Laki**, Attila Novák, and Borbála Siklósi. 2013. “Improved Hungarian Morphological Disambiguation with Tagger Combination.” In: *Text, Speech, and Dialogue*, 8082:280–287. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg.

7. Essentially related publications to the dissertation

[1] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, "Moses: Open Source Toolkit for Statistical Machine Translation," in *Proceedings of the ACL 2007 Demo and Poster Sessions*, Prague, Czech Republic, 2007, pp. 177–180.

[2] A. Novák, "What is good Humor like?," in *I. Magyar Számítógépes Nyelvészeti Konferencia*, Szeged, 2003, pp. 138–144.

[3] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, Philadelphia, USA, 2002, pp. 311–318.

[4] T. Brants, "TnT - A Statistical Part-of-Speech Tagger," in *Proceedings of the Sixth Applied Natural Language Processing (ANLP-2000)*, Seattle, USA, 2000, pp. 224–232.

[5] P. Halácsy, A. Kornai, and C. Oravecz, "HunPos: An open source trigram tagger.," in *Proceedings of the 45th Annual Meeting of the ACL*, Prague, Czech Republic, 2007, pp. 209–212.

[6] G. Orosz and A. Novák, "PurePos 2.0: a hybrid tool for morphological disambiguation," in *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, Hussal, Bulgaria, 2013, pp. 539–545.

[7] K. Toutanova and C. D. Manning, "Enriching the knowledge sources used in a maximum entropy part-of-speech tagger," in *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics - Volume 13*, Hong Kong, China, 2000, pp. 63–70.

[8] A. Søgaard, "Simple semi-supervised training of part-of-speech taggers," in *Proceedings of the ACL 2010 Conference Short Papers*, Uppsala, Sweden, 2010, pp. 205–208.

[9] R. Yeniterzi and K. Oflazer, "Syntax-to-morphology mapping in factored phrase-based statistical machine translation from English to Turkish," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden, 2010, pp. 454–464.

[10] P. F. Brown, V. J. D. Pietra, S. A. D. Pietra, and R. L. Mercer, "The Mathematics of Statistical Machine Translation: Parameter Estimation," *Comput. Linguist.*, vol. 19, no. 2, pp. 263–311, Jun. 1993.

[11] P. Koehn, *Statistical Machine Translation*, 1st ed. New York, NY, USA: Cambridge University Press, 2010.

[12] G. G. Mora and J. A. S. Peiró, "Part-of-Speech Tagging Based on Machine Translation Techniques," in *Proceedings of the 3rd Iberian conference on Pattern Recognition and Image Analysis, Part I*, Girona, Spain, 2007, pp. 257–264.

[13] I. D. El-Kahlout and K. Oflazer, "Exploiting morphology and local word reordering in English-to-Turkish phrase-based statistical machine translation," *Audio Speech Lang. Process. IEEE Trans. On*, vol. 18, no. 6, pp. 1313–1322, 2010.