BASELINE FOR MACEDONIAN STATISTICAL MACHINE TRANSLATION SYSTEM

Miloš Stolić
Institute of Informatics,
Faculty of Natural Sciences and Mathematics
Skopje, Macedonia

ABSTRACT

Statistical machine translation is one of the most widely used machine translation methods. It is able to learn translation patterns from large amounts of language training data for translating from one language to another with or without a human supervisor. This paper describes the necessary tools, data and procedures for the implementation of a statistical machine translation system between Macedonian and English language. An overview of existing linguistic resources for Macedonian language is presented, along with potential further use.

I. INTRODUCTION

Statistical machine translation (SMT) is able to learn translation patterns from large amounts of language training data for translating from one language to another with or without a human supervisor. As such, they are often implemented in research laboratories and commercial systems around the world.

Macedonian language is a South Slavic language, spoken by approximately 2 million speakers. Unfortunately, linguistic resources for Macedonian are rather limited.

On the other hand, English language is one of the most widespread languages in the world, with more than 400 million native speakers. Most computer linguistic research is focused on English, so large amounts of data are freely available, including parallel and monolingual corpora, dictionaries, pre-trained part of speech taggers, semantic parsers, entity recognisers etc.

In this paper, a baseline for creating a statistical machine translation system between English and Macedonian language is given. All available data and tools are available under a free license. Since linguistic resources for English are widely available, only linguistic resources for the Macedonian language are presented, as well as initial results on sentence alignment and part of speech tagging. This data can be used in the training processes of a SMT system.

II. STATISTICAL MACHINE TRANSLATION

The main concept behind SMT is to train a system by aligning large amounts of source language text data with target language text data, thus learning which words or phrases from the target language are most likely to occur given particular words in the source text. The trained system can translate input text by applying statistical rules learned in the training process. The quality of the translation gradually improves by enlarging the training set with new text data.

Moses [1] is the current state of the art SMT decoder. It supports traditional phrase base models, where small chunks of text (phrases) are mapped between two parallel texts, as well as factored models, which expand phrase based models with word level annotation, including lemma, surface form, word class, morphology etc. Factored models are immune to many common errors found in phrase based models, like translation of a noun as a verb, and can overcome data sparseness problems caused by limited training data. However, due to the increase in data which needs to be saved and searched, systems based on factored models tend to be slower during the translation process.

The first step in building a SMT system is obtaining large enough parallel corpus between the source and target language. The corpus needs to be at first sentence aligned, and then word aligned. Factored models also require both source and target corpora to be linguistically annotated.

III. LINGUISTIC DATA

Currently, there are two available parallel Macedonian-English text corpuses – George Orwell's "1984", part of the MULTEXT-East project [2], and a text corpus generated from the Southeast European Times online newspaper. Both text corpuses are encoded in TEI P5 XML [3], where the smallest unit of division is a word. The alignment information is encoded in a separate document shown in Figure 1, containing references to sentence ID's, as specified by the cesAlign DTD, an application of the Corpus Encoding Standard [4].

```
<linkList id="SETmken">
  <linkGrp id="SETmken.1" type="body"
targType="s" >
  <link xtargets="SETmk.1.1 ; SETen.1.1"
certainty="0.864308"/> <!--1:1-->
  <link xtargets="SETmk.1.2 ; SETen.1.2"
certainty="1.21711"/>  <!--1:1-->
```

Figure 1: Parallel Macedonian-English alignment from SET.

A. MULTEXT-EAST

MULTEXT-East (Multilingual Text Tools and Corpora for Central and Eastern European Languages) is a project for developing language resources for east European languages. It is a spin-off of the MULTEXT project, and was EU funded until 1997.

Multext V3 provides linguistic resources for Bulgarian, Croatian, Czech, English, Estonian, Hungarian, Lithuanian, Resian, Romanian, Russian, Slovene, and Serbian language.

Macedonian language has been included into MULTEXT V4. The dataset includes morphosyntactic descriptions (MSD's), word-form lexica [5] as well as an annotated version of George Orwell's "1984" [6].

All translations of "1984" included in the MULTEXT-East dataset are sentence-aligned to the English version as a hub language. The Macedonian and English corpus were sentence aligned using the Vanilla aligner [7]. Additional manual corrections on the alignment have also been performed. The size of the corpus is shown in Table 1.

Table 1: Size of Orwell's "1984" corpus.

Language	Macedonian	English
Sentences	6.821	6.700
Total number of words	95.954	103.997

B. The Southeast European Times

The Southeast European Times (SET) is a Web site sponsored by the US Department of Defence. It covers news and information from South Eastern Europe and it has been published since 2002. All information on the site is released in public domain and can be copied and distributed without permission. The articles are translated into 10 languages: Albanian, Bosnian, Bulgarian, Croatian, English, Greek, Macedonian, Romanian, Serbian and Turkish.

Table 2: Size of the SETimes corpus.

Language	Macedonian	English
Sentences	165k	160k
Total number of tokens	3.6M	3.5M
Total number of words	3.0M	2.9M

A parallel Macedonian-Serbian corpus from SET was described in [8]. By using the same method, a Macedonian-English parallel corpus was generated. The sentence alignment was done using the Hunalign aligner [9] and converted in XML, as pictured in Figure 2. The size of this corpus is available in Table 2.

```
<s xml:id="SETmk.15732.2">БЕЛГРАД , Србија - Полицијата во средата ( 6-ти мај ) го уапси таткото на австралиската тенисерка , родена во Србија , Јелена Докиќ , поради наводната закана дека ќе ја нападне Австралиската амбадсада во Белград со лансирна рампа за проектили . </s>
```

```
<s xml:id="SETen.15732.2">BELGRADE , Serbia - Police arrested on Wednesday ( May 6th ) the father of Serbian-born Australian tennis player Jelena Dokic for allegedly threatening to attack the Australian Embassy in Belgrade with a rocket launcher . </s>
```

Figure 2: Parallel Macedonian-English text from SETimes.

IV. PART OF SPEECH TAGGING

A Part of Speech (PoS) tagged corpus is essential in the creation of a factored model. These additional linguistic annotations are not used in phrase based models.

The morphosyntactic descriptions in MULTEXT are compact string representations of simplified kind of feature structures. The first letter of a MSD encodes a part of speech (grammatical category), e.g. Noun or Verb. The following characters describe attributes of that particular PoS. For instance, the MSD Npfsnn expands to Noun; Type: proper; Gender: feminine; Number: singular; Case: nominative; and Definiteness: no. Table 3 shows all possible attributes for the grammar category "Nouns", specific for the Macedonian language.

Table 3: Attributes and their possible values of the grammatical category Nouns

P	Attribute	Value
1	Type	common
		proper
2	Gender	masculine
		feminine
		neuter
3	Number	singular
		plural
		count
4	Case	nominative
		vocative
		oblique
5	Definiteness	no
		yes
		distal
		proximal

The text corpus from George Orwell's novel "1984" is manually annotated, but not disambiguated. In [6], the performance of T'n'T, a probabilistic part of speech tagger [10], was evaluated. The tagger achieved 83.2% accuracy on unknown words, and 100% accuracy on known words, based on 10-fold cross-validation.

Table 4: PoS tagger accuracy

PoS tagger	Fine-grained accuracy	Coarse-grained accuracy
SVMTool	70.548 %	85.103 %
TreeTagger	72.260 %	85.959 %
T'n'T	63.527 %	77.397 %
MBT	70.205 %	84.418 %

On the other hand, the SET corpus is not annotated. Experiments for projecting PoS tags on the SET corpus using automatic PoS taggers trained on the "1984" corpus have been done [11]. Four PoS taggers have been tested – TreeTagger [12], SVM Tool [13], Memory Based Tagger [14] and the previously mentioned, T'n'T. When using fine-grained annotation, the taggers show 63-70% accuracy. By limiting the PoS tags to coarse categories – i.e. only grammar

category, the accuracy improves to 77-86%, as shown in Table 4.

These results should be taken with caution, since half of the word forms found in SET are not present in the MULTEXT lexica or corpora. Developing a morphological analyzer to handle unknown words would significantly improve accuracy.

V. FACTORED TRANSLATION MODEL

A factored translation model, unlike phrase translation models, operates on more general word representations, such as lemmas instead of surface forms of words. An input word is actually complex data, consisted of different factors, like surface form, lemma and part of speech. The translation of such factored input is consisted of a sequence of mappings. First, an output lemma is mapped for the given input lemma. Then, input and output PoS are mapped. Finally, surface forms are generated based on the previous output lemmas and PoS. Since multiple choices are available for each step, each input phrase can be expanded into a list of translation options.

VI. TRAINING

The first step in the training process is finding word alignments using GIZA++ [15]. It is the most popular tool for statistical word alignment, which trains translation models from parallel training data by implementing IBM Models 1-5. GIZA++ aligns words and phrases in both Macedonian-English and English-Macedonian direction. Since the source and target text are annotated, the word alignment can be between surface forms or lemmas. A translation model is acquired by extracting phrase mappings over factored representations from the alignments. The set of phrase mappings is scored based on relative counts and word-based translation probabilities.

Along with the factored translation model, an n-gram language model over the surface forms of the target language is also trained. Such models can be built with the proprietary SRI Language Modelling Toolkit [16] or the open source IRST Language Modelling Toolkit [17] and RandLM [18]. This language model is used to calculate the probability that the translation result from the translation model is a valid phrase or sentence in the target language. Then, the language model, mapping steps and generation are combined in a log-linear model.

VII. CONCLUSION AND FURTHER WORK

Even though there are limited linguistic resources for the Macedonian language, building a statistical machine translation system is more than possible. It would be interesting to see how such system would perform compared to commercially available software, like Google Translate, as well as traditional phrase based models. Developing new Macedonian-English parallel corpora is of utmost priority, along with large dictionary and lexicon.

REFERENCES

[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, E. Herbst, "Moses: Open Source Toolkit for Statistical Machine Translation", in *Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session*, Prague, Czech Republic, June 2007.

[2] T. Erjavec, "MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora", in *Fourth International Conference on Language Resources and Evaluation*, LREC-04, Paris, 2004.

[3] TEI Consortium, "Guidelines for Electronic Text Encoding and Interchange", http://www.tei-c.org/P5/

[4] N. Ide, "Corpus Encoding Standard: SGML guidelines for encoding linguistic corpora", in *First International Conference on Language Resources and Evaluation*, LREC'98, Granada, ELRA, 1998.

[5] K. Zdravkova, A. Ivanovska, S. Džeroski, T. Erjavec, "Learning Rules for Morphological Analysis and Synthesis of Macedonian Nouns", in *Proceedings of IS 2005, the 8th International Multiconference on the Information Society*, 11-17 October 2005, Ljubljana, pp. 195-198.

[6] V. Vojnovski, S. Dzeroski, T. Erjavec, "Learning POS tagging from a tagged Macedonian text corpus", in *Proceedings of IS 2005, the 8th International Multiconference on Information Society*, 11-17 October 2005, Ljubljana, pp. 199-202.

[7] W. Gale and K. Church, "Program for aligning sentences in bilingual corpora", in *Computational Linguistics* 19, pp. 75-102, 1993.

[8] M. Stolić and K. Zdravkova, "Resources for Machine Translation of the Macedonian Language", in *ICT-Innovations Conference,* 2009.

[9] D. Varga, L. Németh, P. Halácsy, A. Kornai, V. Trón, V. Nagy, "Parallel corpora for medium density languages", in *Proceedings of the RANLP 2005 Conference*, pp. 590-596.

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

[11] M. Stolić, A. Kuzmanovska and K. Zdravkova, "Evaluating Part of Speech Taggers on the Macedonian Language", unpublished.

[12] H. Schmid, Probabilistic "Part-of-Speech Tagging Using Decision Trees", 1994.

[13] J. Gimenez and L. Marquez, "SVMTool: A general POS tagger generator based on Support Vector Machines", 2004.

[14] W. Daelemans, J. Zavrel, P. Berck and S. Gillis, "MBT: A Memory-Based Part of Speech Tagger-Generator", in Proceedings of Fourth Workshop on Very Large Corpora, 1996, pp 14-27.

[15] F. Och, H. Ney, "A Systematic Comparison of Various Statistical Alignment Models", in *Computational Linguistics*, volume 29, number 1, pages 19-51.

[16] A. Stolcke, "SRILM - An Extensible Language Modeling Toolkit", in *Proceedings of International Conference Spoken Language Processing*, Denver, Colorado, September 2002.

[17] M.Federico, N. Bertoldi and M. Cettolo, "IRSTLM: an Open Source Toolkit for Handling Large Scale Language Models", in *Proceedings of the Interspeech 2008*, Brisbane, Australia, 2008.

[18] D. Talbot and M. Osborne, "Randomised Language Modelling for Statistical Machine Translation", ACL, Prague, Czech Republic 2007.