TED Polish-to-English translation system for the IWSLT 2012

Krzysztof Marasek

Multimedia Department
Polish-Japanese Institute of Information Technology, Warsaw, Poland
kmarasek@pjwstk.edu.pl

Abstract

This paper presents efforts in preparation of the Polish-to-English SMT system for the TED lectures domain that is to be evaluated during the IWSLT 2012 Conference. Our attempts cover systems which use stems and morphological information on Polish words (using two different tools) and stems and POS.

1. Introduction

Polish, one of the West-Slavic languages [1], due to its complex inflection and free word order, forms a challenge for statistical machine translation (SMT). Polish grammar is quite complex: seven cases, three genders, animate and inanimate nouns, adjectives agreed with nouns in terms of gender, case and number and a lot of words borrowed from other languages which are often inflected similarly to those of Polish origin. These cause problems in establishing vocabularies of manageable sizes for translation to/from other languages and sparseness of data for statistical model training. Despite of ca. 60 millions of Polish speakers worldwide the number of publicly available resources for the preparation of SMT systems is rather limited, thus progress in that domain is slower than for other languages. In this paper, our efforts in preparation of the Polish-to-English SMT system for the TED task, part of the IWSLT 2012 evaluation campaign, MT optional track, are described.

The remainder of the paper is structured as follows. In section 2 Polish data preparation is described, section 3 deals with English, 4 with training of the translation and language models, and section 5 presents our results. Finally, the paper concludes with a discussion about encountered issues and future perspectives in sections 6 and 7.

2. Polish data preparation

Training, development and evaluation data consists of the Polish translation of TED lectures and its English origin. This has been prepared by FBK [2]. The available data set consists of ca. 2.27 millions of untokenized words on the target side. The transcripts are given as pure text (UTF-8 encoding), one or more sentences per line, and are aligned at language pair level. The organizers also provide a lot of monolingual data (English) and the PL-EN Europarl v.7 parallel corpus.

Some manual preprocessing of training data was necessary.

After extracting the transcripts from the supplied XML files the same number of lines for both languages were obtained, but with some discrepancies in the parallel text. Those differences were caused mostly by repetitions in the Polish text and some additional remarks (like “Applause” or “Thanks”) which were not present in the English text. 28 lines had to be manually corrected for the whole set of 134325 lines. Without trying to judge the TED data translation quality, but as a Polish native speaker, it left an impression that, at least part of the talks were translated by volunteers, making the training material a bit noisy. Moreover, a lot of English proper names are inserted into Polish text.

The vocabulary sizes (extracted using SRILM [3]) were 198622 for Polish and 91479 for English, which exposes the fundamental problem for the translation – the huge difference in the vocabulary sizes.

Tokenization of input data was done using standard tools delivered with Moses [4], with an extension created by FBK for Polish.

Before a translation model was trained, the usual preprocessing was applied, such as removing long sentences (threshold 60) and sentences with length difference exceeding a certain threshold. This was done again using scripts from the Moses toolkit.

The final tokenized, lowercased and cleaned training corpus for Polish and English was 132307 lines long, but with an even greater difference in vocabulary sizes – 47250 for English vs. 123853 for Polish.

This large difference between source and target vocabulary sizes shows the necessity of using additional knowledge sources. Initially, we decided to limit the size of the Polish vocabulary by using stems instead of surface forms. Following that, we tried using morphosyntactic tagging as an additional source of information for the SMT system.

2.1. Stems extraction for Polish

Inspired by the works of Bojar [6], we tried to use stems of Polish words instead of its surface forms with the purpose of reducing the vocabulary size difference. Since the target language is English, it was not necessary to build models which will convert stems to correct grammatical forms – the target was a normal English sentence (surface forms).

For that purpose, a set of freely available tools prepared by the NLP group of the Wrocław Technical University was used. This set of NLP-tools (http://nlp.pwr.wroc.pl) can be used to perform the following tasks:

- Tokenisation — division into tokens and sentences
- Morphosyntactic analysis using the available analysers and dictionaries (including Morfeusz SGJP/SIAT), but also user-supplied dictionaries
- Morphosyntactic tagging
- Shallow parsing (understood as chunking)
- Turning running text into a sequence of feature vectors (using WCCL formalism, useful for further NLP tasks)

From this, two main components were used:

- MACA [8] – a universal framework to join different sources of morphological information, including the existing resources as well as user-provided dictionaries. This framework allows writing simple configuration files that define tokenisation strategies and the behavior of morphological analysers, including simple tagset conversion.

- WCRFT [7] – morphosyntactic tagger which brings together Conditional Random Fields and tiered tagging (where grammatical information is split into several tiers, usually one tier is used for each of grammatical classes).

The tools, when used in a sequence, form XML-formatted output containing for each token: its surface form, stem and morphosyntactic tag (tags).

If stems are only taken from the Polish TED training data, the vocabulary (for data cleaned as previously) is substantially reduced to only 44102 words.

2.2. Morphosynactic tagging: Wrocław tools

The tagset used by the Wrocław's analyzers could have been changed, but it was most straightforward to use the standard settings, where the IPIC (IPI PAN Corpus, Polish National Corpus [9]) tagset is used. This particular tagset allows for much more fine-grained tagging compared to traditional parts-of-speech. Each tag contains a grammatical class and zero or more values for certain attributes. Each grammatical class defines a set of attributes whose values must be specified. For instance, nouns require that number, gender and case attributes are specified, and adverbs require the degree attribute. This in turn causes specific segmentation of input text, where some words are split into several tokens, thus tokenization differs from the one delivered by standard Moses tools. This causes some problems when building parallel corpora. In order to avoid these problems, additional markers were placed at the end of each input line.

The tagger tries to disambiguate the grammatical forms giving the set of most probable tags. Usually, just one tag is provided and only in really undistinguishable cases all possible tags are given, as in the following example (pl.gen. *man* from sin.nom. *man* or pl.nom *people*):

```
<tok>
<orth>ludzi</orth>
<lex     disamb="1">     <base>człowiek</base>
<ctag>subst:pl:gen:m1</ctag></lex>
<lex      disamb="1">      <base>ludzie</base>
<ctag>subst:pl:gen:m1</ctag></lex>
</tok>
```

In such a case only the first form (first stem) was taken for further processing.

2.3. Morphosynactic tagging: our tools

In several projects related to speech technology a grave demand for text normalization is observed. Text normalization is the process of converting any abbreviations, numbers and special symbols into corresponding word sequences. In particular, normalization is responsible for:

1. expansion of abbreviations in the text into their full form;
2. expansion of any numbers (e.g. Arabic, Roman, fractions) into their appropriate spoken form;
3. expansion of various forms of dates, hours, enumerations and articles in contracts and legal documents into their proper word sequences.

This task, although seemingly simple, is in fact quite complicated – especially in languages like Polish which has 7 cases and 15 gender forms for nouns and adjectives, with additional dimensions for other word classes. That is why most abbreviations have multiple possible expansions and each number notation over a dozen outcomes.

To solve this task we prepared tools [10] which we also try to use for morphosyntactic tagging of Polish texts.

The system consists of a decoder, a language model and a set of expansion rules. The expansion rules are used in the expansion of commonly used abbreviations and written date and number forms. A synchronous Viterbi style decoder that generates a list of hypotheses ordered by the values retrieved from the language model is used. Each time the text contains a word sequence that could be expanded; all the possible expansions are fed into the decoder. Because the expansion of long numbers or some abbreviations expects that several words need to be added at once, hypotheses of varying lengths may end up competing against each other. This is remedied by the normalization of hypotheses' probabilities to their lengths. Such normalization is equivalent to the addition of a heuristic component commonly used in asynchronous decoders like A∗. The language model itself is a combination of three models with a range of n=3 for the individual words, n=5 for word stems and n=7 for grammatical classes. The Evolution Strategy ($\mu + \lambda$) is used for optimization of model weights, especially:

1. weights of 30 text domain sets (10 parameters for each model),
2. linear interpolation weight for all n-grams in all models. The weights depended on the frequency of occurrence of given n-gram - there were 5 ranges of frequency,
3. linear interpolation weights for the word, stems and grammar classes models (combining the smaller models into one larger), with perplexity of the final model on development set as a quality criterion.

The outcome of the system is also a morphosyntactic tagging of tokens, however no disambiguation is done. Instead, a numerical value describing all possible tags for a given form is stored, eg.:

```
id = 15
features:
adj;acc;sg;m_os;;pos;;
adj;acc;sg;m_zyw;;pos;;
adj;gen;sg;m_nie_zyw;;pos;;
adj;gen;sg;m_os;;pos;;
adj;gen;sg;m_zyw;;pos;;
adj;gen;sg;neu;;pos;;
```

for the surface form "tego" (stem: "ten", eng. *this*).

It should be also noted, that stems are generated only for words from a given vocabulary (for other words OOV symbol is placed) and proper names, foreign words, spellings and abbreviations are recognized and special symbols are inserted instead of stems as in following example:

```
plan|plan|5 był|być|106 w|*letter|0
pełni|pełnia|9 gotowy|gotowy|18 w|*letter|0
dziewięćdziesiątym|dziewięćdziesiąty|255
ósmym|ósmy|255 roku|rok|93  nosił|nosić|106
nazwę|nazwa|10 digital|oov|-2 Millennium|OOV|-
2 Copyright|OOV|-2 act|OOV|-2 .|.|
```

Our tool uses Windows-1250 Eastern Europe character encoding, thus it was necessary to convert data from/to UTF-8 encoding used by all other tools. The decoding procedure showed several UTF-8 special characters used in the original text (like musical notes, etc.) which added some manual work to remove those unnecessary symbols.

3. English data preparation

Preparation of English data was less complicated. For the baseline (surface form) and stems of Polish, only surface forms of English TED data was used. For the factored model, English text was tagged using Stanford CoreNLP tools [11,12]. Stanford CoreNLP integrates all necessary NLP tools, including the parts-of-speech (POS) tagger and provides model files for analysis of English, providing the base forms of words, their parts of speech, recognition of named entities, normalization of dates, times, and numeric quantities, and marks of the structure of sentences in terms of phrases.

4. Training and tuning procedure

Only in-domain data for training of the SMT system was used, mainly because of our lack of experience in translation model adaptation. Also, no other English data for language modeling was used. The supplied Euro-parlament data was from a too distant domain and our attempts to use Google n-grams ended without success (noisy data, tools which we have did not work properly on such huge large data sets). TED talks corpus consists of data which varies significantly with respect to the topics or domain, but has a rather homogeneous presentation style. Moreover, the TED training data perfectly matches the test condition, so we assume that the possible gain from using other data could be limited. It was also our intention to focus our work on researching proper factors combination and configuration of the SMT training.

Thus, TED lectures data [2] was used for training in 4 main modes:

BASE Polish surface form to English surface form

STEM Polish stems to English surface form

FCT1 Polish factors (surface form | stem | extended morphosytactic tag from Wrocław tools) to English factors (surface form | stem | POS from Stanford CoreNLP),

FCT2 Polish factors (surface form | stem | numerical morphosytactic tag from our tool) to English factors (surface form | stem | POS from Stanford CoreNLP).

As development and evaluation data again TED talks are used [2]. The set "iwslt2012-dev2010" consists of 767 lines. Testing of the system was done on "iwslt2012-tst2010" set build of 1564 lines. All development and test data has been prepared for all 4 modes of the SMT training.

All the language models used are 5-gram interpolated language models with Kneser-Ney discounting and were trained with the SRILM toolkit [12]. This includes also language models trained on stems and grammatical tags.

The word alignment of the parallel corpora was generated using the GIZA++-Toolkit [5]. Afterwards, the alignments were combined using the grow-diag-final-and heuristic. The phrases were extracted and scored using the Moses toolkit [4]. For the BASE, FCT1 and FCT2 systems several reordering models were tested. Only marginal improvement on test data was achieved compared to the standard setting "msd-bidirectional-fe".

Tuning was done using MERT Moses' implementation [14] on development data. New weights were then used for testing. A lot of work was spent on finding good composition of factors for translation, generation and decoding steps of the factored models. However, as shown in the next section, we did not find efficient factors yet.

5. Evaluation

For training all the data has been lowercased and tokenized. The evaluation needs data to be recased to its original form. For that, a model was trained using standard Moses tool train-recaser.pl. Evaluation results are presented in Tables 1 and 2.

Table 1: Results of the evaluation, truecase and punctation

TASK	SYSTEM	BLEU	METEOR	WER	PER	TER	GTM	NIST
	BASE	0.2	0.56	0.66	0.52	61.42	0.55	5.64
dev2010	STEM	0.19	0.56	0.66	0.54	62.41	0.53	5.43
	FCT1	0.13	0.47	0.64	0.57	61.88	0.5	4.23
	FCT2	0.1						2.96
	BASE	0.15	0.49	0.74	0.59	69.04	0.49	4.9
tst2010	STEM	0.14	0.49	0.73	0.6	69.21	0.48	4.77
	FCT1	0.11	0.43	0.69	0.6	66.15	0.46	3.92
	FCT2	0.09						2.71
	BASE	0.19	0.54	0.68	0.55	64.19	0.53	5.44
tst2011	STEM	0.17	0.54	0.69	0.57	65.07	0.51	5.2
	FCT1	0.14	0.47	0.64	0.57	61.84	0.49	4.39
	FCT2							
	BASE	0.15	0.48	0.72	0.6	67.96	0.48	4.98
tst2012	STEM	0.14	0.48	0.72	0.6	68.31	0.47	4.78
	FCT1	0.11	0.42	0.69	0.62	66.14	0.45	3.6

Table 2: Results of the evaluation, no casing and no punctation

TASK	SYSTEM	BLEU	METEOR	WER	PER	TER	GTM	NIST
	BASE	0.19	0.53	0.67	0.54	64.46	0.53	5.78
dev2010	STEM	0.17	0.53	0.68	0.56	65.82	0.51	5.5
	FCT1	0.13	0.45	0.66	0.58	64.97	0.48	4.33
	FCT2	0.1						2.88
	BASE	0.14	0.46	0.76	0.62	73.12	0.47	5.05
tst2010	STEM	0.13	0.46	0.76	0.63	73.66	0.45	4.86
	FCT1	0.11	0.41	0.72	0.62	70.05	0.44	4.09
	FCT2	0.08						2.67
	BASE	0.18	0.5	0.7	0.57	67.44	0.51	5.64
tst2011	STEM	0.16	0.5	0.71	0.59	69.19	0.49	5.33
	FCT1	0.13	0.44	0.67	0.59	65.64	0.47	4.48
	FCT2							
	BASE	0.14	0.44	0.74	0.61	71.53	0.46	5.13
tst2012	STEM	0.13	0.44	0.74	0.63	72.52	0.44	4.85
	FCT1	0.1	0.39	0.72	0.64	70.51	0.43	3.61

TASK describes the test set, SYSTEM is one of the systems described in section 4, and BLEU, METEOR, WER, PER, TER, GTM and NIST are appropriate evaluation scores (see en.wikipedia.org/wiki/Evaluation_of_machine_translation for explanation). For the BASE, STEM, FCT1systems the scoring was done by the IWSLT evaluation team [17], for the system FCT2 scoring was done in house using mteval-v12 NIST script for dev2010 and tst2010 datasets only.

6. Discussion

As mentioned in section 4, a lot of work was spent trying to find the best combination of factors for translation, generation and decoding steps within the Moses framework. Unfortunately, a lot of combination ended with decoder errors, with no clear reasons given. This showed that more experience to use those advanced features is definitely needed.

Many researchers claim that word alignment is crucial for good SMT results. The recent study of Wróblewska [15] shows that, in her experiments, best precision of word alignment was achieved if the Polish side of the parallel corpus was lemmatized. This reduces the number of items in the lemma dictionary and approximates the English token dictionary. She does not give an answer to whether lemmatising the English part of the parallel corpus is necessary. Her results somewhat resemble the work presented in this paper.

It also clear that TED talks is a difficult task, at least on the Polish side (huge vocabulary, many long lines). Just for comparison, on the BTEC corpus [16] we obtained better results (NIST=14.27 BLEU=0.89 on development set using mteval-v12 script). It is because BTEC consists of short, clear sentences without any foreign terms (usually inflected in Polish) as it is in the TED talks.

7. Conclusions

The conducted experiments are only a first step towards building the final Polish-to-English SMT system. We tried to use surface forms, stems and two kinds of factors describing grammatical properties of Polish words and surface forms, stems and POS for English. In the near future, we will try to use more data (Europarl) for the SMT preparation and optimize the system for the in-domain data. In further research, we would like to investigate the usage of surface forms and stems simultaneously on the Polish side and look more deeply into works done for other Slavic languages.

8. Acknowledgements

This work is sponsored by the EU-Bridge 7 FR project (grant agreement no. 287658) and statutory works of the PJIIT (ST/MUL/4/2011).

9. References

[1] Jagodziński G., “A Grammar of Polish Language”, http://grzegorj.w.interia.pl/gram/en/gram00.html

[2] Cettolo M, Girardi C., Federico M., “WIT3: Web Inventory of Transcribed and Translated Talks”. In *Proc. of EAMT*, pp. 261-268, Trento, Italy, 2012

[3] Stolcke A., "SRILM - An Extensible Language Modeling Toolkit", in *Proc. Intl. Conf. Spoken Language Processing*, Denver, Colorado, September 2002

[4] Koehn P., Hoang H., Birch A., Callison-Burch C., Federico M., Bertoldi N., Cowan B., Shen W., Moran C., Zens R., Dyer R., Bojar O., Constantin A., and Herbst E., “Moses: Open Source Toolkit for Statistical Machine Translation,” in *Proceedings of ACL 2007*, Demonstration Session, Prague, Czech Republic, 2007.

[5] F. J. Och and H. Ney, “A systematic comparison of various statistical alignment models,” *Computational Linguistics*, vol. 29, no. 1, pp. 19–51, 2003.

[6] Bojar O., “Rich Morphology and What Can We Expect from Hybrid Approaches to MT”. *Invited talk at International Workshop on Using Linguistic Information for Hybrid Machine Translation* (LIHMT-2011), http://ufal.mff.cuni.cz/~bojar/publications/2011-FILE-bojar_lihmt_2011_pres-PRESENTED.pdf , 2011

[7] Radziszewski A., “A tiered CRF tagger for Polish”, in: *Intelligent Tools for Building a Scientific Information Platform: Advanced Architectures and Solutions*, editors: Membenik R., Skonieczny L., Rybiński H., Kryszkiewicz M., Niezgódka M., Springer Verlag, 2013 (to appear)

[8] Radziszewski A., Śniatowski T., “Maca: a configurable tool to integrate Polish morphological data”, *Proceedings of the Second International Workshop on Free/Open-Source Rule-Based Machine Translation*, FreeRBMT11, Barcelona, 2011

[9] Przepiórkowski A., Bałko M., Górski R., Lewandowska-Tomaszczyk B., „*Narodowy Korpus Języka Polskiego*”, PWN Warszawa, 2012

[10] Brocki Ł., Marasek K., Korzinek D., “Multiple Model Text Normalization for the Polish Language”, *The 20th International Symposium on Methodologies for Intelligent Systems ISMIS-2012*, Macau, 4-7 December 2012 (in press)

[11] Toutanova K, Klein D., Manning Ch., and Singer Y., “Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network”, in *Proceedings of HLT-NAACL* 2003, pp. 252-259.

[12] Finkel J., Grenager T., and Manning Ch., Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics* (ACL 2005), pp. 363-370.

[13] Stolcke A., “SRILM – An Extensible Language Modeling Toolkit”, *International Conference on Spoken Language Processing*, Denver, Colorado, USA, 2002.

[14] Bertoldi N., Haddow B., Fouet J.-B., “Improved Minimum Error Rate Training in Moses”, *The Prague Bulletin of Mathematical Linguistics*, February 2009, pp.1-11

[15] Wróblewska A., “Polish-English word alignment: preliminary study”, in Ryżko D., Rybiński H., Gawrysiak M., Kryszkiewicz M, editors, *Emerging Intelligent Technologies in Industry, volume 369 of Studies in Computational Intelligence*, pp. 123–132, Springer-Verlag, Berlin, 2011.

[16] Takezawa T., Kikui G., Mizushima M., Sumita E., “Multilingual Spoken Language Corpus Development for Communication Research”, *Computational Linguistics and Chinese Language Processing*, Vol. 12, No. 3, September 2007, pp. 303-324

[17] M. Federico, M. Cettolo, L. Bentivogli, M. Paul, S. Stueker, Overview of the IWSLT 2012 Evaluation Campaign, *In Proc. of IWSLT*, Hong Kong, HK, 2012