Data Pre-Processing to Train a Better Lithuanian-English MT System

Daiga Deksne (daiga.deksne@tilde.lv), Raivis Skadiņš (raivis.skadins@tilde.lv)

Abstract

In this paper, we present the results of a series of experiments done to improve the quality of a Lithuanian-English statistical MT (SMT) system. We particularly focus on word alignment and out of vocabulary issues in SMT translating from a morphologically rich language into English.

Introduction

As **Lithuanian is highly inflected language,** the words change the form according to grammatical function. That means that the endings of nouns, pronouns, adjectives, numerals and verbs change depending on certain features. **English instead does not have such a rich feature system**.

This difference between languages **significantly impacts word and phrase alignment** when training an SMT system. Typically one or two forms of an English word have to be aligned to more than ten different surface forms of a corresponding Lithuanian word. Lithuanian verbs have prefixes indicating negation and other semantic features while English verbs do not have prefixes and such information is expressed using modifying words.

Many word forms are not as common as others in the corpus, therefore a **Lithuanian-English SMT system does not translate all word forms equally well**.

It is very **common** to get **many out of vocabulary words** when translating from Lithuanian into English.

Chosen approach

- Four different experiments using the DGT-TM parallel corpus* where performed to train the MT system which performs better also on not so common word forms
- Results where compared to the baseline SMT system trained on the original DGT corpus without any data pre-processing
- After result evaluation another experiment was performed on a larger, more general corpus training a baseline system and a system with a pre-processed data using the best method from previous experiments
- All systems were trained using the LetsMT! platform which is based on the Moses SMT toolkit
- In all the experiments the corpus was transformed using **finite state transducers**

* http://langtech.jrc.it/DGT-TM.html
(ca. 806 K parallel and 703 K monolingual sentences)

Experiments

Prefixes and endings as separate tokens
System #1

- We apply several transformation rules to a Lithuanian text corpus
- If possible, endings and prefixes are separated from word stems
- A list of non-inflected part of speech words is included in transducer. They are not changed
- A list of prefixes and a list of endings are included in the transducer
- Combination of an optional prefix, a stem and an optional ending forms the word
- The stem can be any sequence of letters which is at least two symbols long
- The transformed word is in form:

prefix- stem -ending

Original sentence:

Priedas ir Protokolas yra neatskiriama šio Susitarimo dalis.

Pre-processed sentence:

Pried -as ir Protokol -as yr -a ne- atskiriam -a ši -o Susitar -imo dal -is.

Prefixes separated, endings replaced by tense and number feature values
System #2

- Prefixes are separated from word stems
- Endings are replaced by tense and number feature values
- Transformed word is in form:

prefix- stem&featurevalues

- The same ending can symbolize several feature values. Some examples of feature value tags – PRESPAST, SG, PL, PLPRES
- If some particular ending can be both – singular and plural form ending – number feature value is not used
- A list of non-inflected part of speech words and a list of personal pronouns are included in transducer, they are not changed
- There is no distinction between verb stems and other stems in transducer. This leads to situation when tense feature values are also assigned to noun stems

Pre-processed sentence:

Pried&PRES ir Protokol&PRES yr&PRES ne- atskiriam&PRES ši&PRESPAST Susitar&SGPRES dal&PRES.

	Singular	Plural
Nominative	PRES (as)	PRESPAST (ai)
Genitive	PRESPAST (o)	- (ų)
Dative	SG (ui)	PL (ams)
Accusative	PRES (ą)	PAST (us)
Instrumental	PRES (u)	PRESPAST (ais)
Locative	SG (e)	PL (uose)

Prefixes separated, all endings replaced by number feature values and verb endings also by time feature values
System #3

- The two lists of endings – the verb endings and the other endings – are used to avoid the drawbacks of System #2
- The tense feature is applied only to verb endings
- Transducer has a full list of verb stems for which the verb endings are allowed
- Other endings are allowed to any two or more letter combination which is not in the verb stem list
- The verb stems are with a higher weight than other stems
- Same as before, a list of non-inflected part of speech words and a list of personal pronouns is used in the transducer

Pre-processed sentence:

Pried& ir Protokol& yr&PRES ne- atskiriam&SG ši&PRESPAST Susitar&SG dal&PRES.

- Comparing with System #2 the situation has improved, the two first paradigm nouns “priedas” and “protokolas” do not have the tense feature values. But the noun stem “dal” still has the tense feature values tag as this stem also can be a verb stem
- Now nouns have tags SG and PL. Some forms do not have number feature values, because they can be either singular or plural for the nouns of the different paradigms

	Singular	Plural
Nominative	- (as)	- (ai)
Genitive	SG (o)	- (ų)
Dative	SG (ui)	PL (ams)
Accusative	SG (ą)	- (us)
Instrumental	- (u)	PL (ais)
Locative	SG (e)	PL (uose)

Prefixes separated, endings deleted
System #4

- Prefixes are separated as in the previous experiments
- But endings are deleted
- The transformed sentence contains only the non-inflected words and the stems of the inflected part of speech words

Pre-processed sentence:

Pried ir Protokol yr ne- atskiriam ši Susitar dal.

Evaluation

All the systems with pre-processed data where evaluated on a random subset of 1,000 sentences from the DGT-TM corpus using BLEU, NIST, TER and METEOR automatic metrics.

System	BLEU	NIST	TER	METEOR
Baseline	49.04	9.2774	48.54	0.4214
System #1	47.53	9.1871	50.02	0.4199
System #2	49.17	9.2546	49.07	0.4208
System #3	**49.22**	**9.2886**	**48.28**	**0.4241**
System #4	47.99	9.1072	49.83	0.4186

We also evaluated all systems on a balanced out-of-domain test corpus (512 sentences).

System	BLEU	NIST
Baseline	15.14	**4.9721**
System #1	14.92	4.9487
System #2	13.68	4.7170
System #3	**15.38**	4.9600
System #4	13.72	4.6859

- System #4 has the lowest BLEU score on both evaluation sets
- Only System #3 exceeds the baseline system

Large-scale experiment

- The results obtained using the DGT-TM corpus show that a better MT system can be built by applying different pre-processing methods
- The next step: train SMT system on a larger, more general corpus (5.3M parallel and 81M monolingual sentences)
- Baseline system without data pre-processing
- The system with data pre-processing as in System #3 (the best results on a smaller corpus)
- The system with pre-processed data outperforms the baseline system by **0.59 BLEU** points

System	BLEU
Baseline	37.83
System with pre-processing	**38.42**

- Human evaluation was also performed to compare both systems
- Nine human evaluators where asked to give preference to translation of the first or of the second system
- The results of the system with pre-processed data are slightly better, in **50.99% (±3.55%)** of cases human evaluators judged its output to be better than the baseline system’s output
- However, evaluation results are not sufficient to say with strong confidence that the system with data pre-processing is better, because the difference between the systems is not statistically significant

Conclusions and Future Work

- Experiments show that it is **possible to improve the quality** of SMT translation from a highly inflected language into English **by pre-processing the training data**
- Even a simple method gives significant improvement in SMT systems trained both on small and a large corpora
- This paper describes only 4 simple ways of data pre-processing
- We are considering the use of more advanced tools such as **part of speech tagger** or **morphological analyzer** instead of finite state transducers with very limited lexicon
- Experiments with different length of prefixes and suffixes are also considered

Acknowledgements

The research within the project LetsMT! leading to these results has received funding from the ICT Policy Support Programme (ICT PSP), Theme 5 – Multilingual web, grant agreement no 250456.

About

www.tilde.com

Daiga Deksne daiga.deksne@tilde.lv

Raivis Skadiņš raivis.skadins@tilde.lv