Morpheme- and POS-based IBM1 scores and language model scores for translation quality estimation

Maja Popović
German Research Center for Artificial Intelligence (DFKI)
Language Technology (LT), Berlin, Germany
maja.popovic@dfki.de

Abstract

We present a method we used for the quality estimation shared task of WMT 2012 involving IBM1 and language model scores calculated on morphemes and POS tags. The IBM1 scores calculated on morphemes and POS-4grams of the source sentence and obtained translation output are shown to be competitive with the classic evaluation metrics for ranking of translation systems. Since these scores do not require any reference translations, they can be used as features for the quality estimation task presenting a connection between the source language and the obtained target language. In addition, target language model scores of morphemes and POS tags are investigated as estimates for the obtained target language quality.

1 Introduction

Automatic quality estimation is a topic of increasing interest in machine translation. Different from evaluation task, quality estimation does not rely on any reference translations – it relies only on information about the input source text, obtained target language text, and translation process. Being a new topic, it still does not have well established baselines, datasets or standard evaluation metrics. The usual approach is to use a set of features which are used to train a classifier in order to assign a prediction score to each sentence.

In this work, we propose a set of features based on the morphological and syntactic properties of involved languages thus abstracting away from word surface particularities (such as vocabulary and domain). This approach is shown to be very useful for evaluation task (Popović, 2011; Popović et al., 2011; Callison-Burch et al., 2011). The features investigated in this work are based on the language model (LM) scores and on the IBM1 lexicon scores (Brown et al., 1993).

The inclusion of IBM1 scores in translation systems has shown experimentally to improve translation quality (Och et al., 2003). They also have been used for confidence estimation for machine translation (Blatz et al., 2003). The IBM1 scores calculated on morphemes and POS-4grams are shown to be competitive with the classic evaluation metrics based on comparison with given reference translations (Popović et al., 2011; Callison-Burch et al., 2011). To the best of our knowledge, these scores have not yet been used for translation quality estimation. The LM scores of words and POS tags are used for quality estimation in previous work (Specia et al., 2009), and in our work we investigate the scores calculated on morphemes and POS tags.

At this point, only preliminary experiments have been carried out in order to determine if the proposed features are promising at all. We did not use any classifier, we used the obtained scores to rank the sentences of a given translation output from the best to the worst. The Spearman's rank correlation coefficients between our ranking and the ranking obtained using human scores are then computed on the provided manually annotated data sets.

2 Morpheme- and POS-based features

A number of features for quality estimation have been already investigated in previous work (Specia et al., 2009). In this paper, we investigate two sets of

Proceedings of the 7th Workshop on Statistical Machine Translation, pages 133–137,
Montréal, Canada, June 7-8, 2012.

features which do not depend on any aspect of translation process but only on the morphological and syntactic structures of the involved languages: the IBM1 scores and the LM scores calculated on morphemes and POS tags. The IBM1 scores describe the correspondences between the structures of the source and the target language, and the LM scores describe the structure of the target language. In addition to the input source text and translated target language hypothesis, a parallel bilingual corpus for the desired language pair and a monolingual corpus for the desired target language are required in order to learn IBM1 and LM probabilities. Appropriate POS taggers and tools for splitting words into morphemes are necessary for each of the languages. The POS tags cannot be only basic but must have all details (e.g. verb tenses, cases, number, gender, etc.).

2.1 IBM1 scores

The IBM1 model is a bag-of-word translation model which gives the sum of all possible alignment probabilities between the words in the source sentence and the words in the target sentence. Brown et al. (1993) defined the IBM1 probability score for a translation pair f_1^J and e_1^I in the following way:

$$P(f_1^J|e_1^I) = \frac{1}{(I+1)^J} \prod_{j=1}^{J} \sum_{i=0}^{I} p(f_j|e_i) \quad (1)$$

where f_1^J is the source language sentence of length J and e_1^I is the target language sentence of length I.

As it is a conditional probability distribution, we investigated both directions as quality scores. In order to avoid frequent confusions about what is the source and what the target language, we defined our scores in the following way:

- source-to-hypothesis (sh) IBM1 score:

$$\text{IBM1}_{sh} = \frac{1}{(H+1)^S} \prod_{j=1}^{S} \sum_{i=0}^{H} p(s_j|h_i) \quad (2)$$

- hypothesis-to-source (hs) IBM1 score:

$$\text{IBM1}_{hs} = \frac{1}{(S+1)^H} \prod_{i=1}^{H} \sum_{j=0}^{S} p(h_i|s_j) \quad (3)$$

where s_j are the units of the original source language sentence, S is the length of this sentence, h_i are the units of the target language hypothesis, and H is the length of this hypothesis.

The units investigated in this work are morphemes and POS-4grams, thus we have the following four IBM1 scores:

- MIBM1$_{sh}$ and MIBM1$_{hs}$:

  IBM1 scores of word morphemes in each direction;

- P4IBM1$_{sh}$ and P4IBM1$_{hs}$:

  IBM1 scores of POS 4grams in each direction.

2.2 Language model scores

The n-gram language model score is defined as:

$$P(e_1^I) = \prod_{i=1}^{I} p(e_i|e_i...e_{i-n}) \quad (4)$$

where e_i is the current target language word and $e_i...e_{i-n}$ is the history, i.e. the preceeding n words.

In this paper, the two following language model scores are explored:

- MLM6:

  morpheme-6gram language model score;

- PLM6:

  POS-6gram language model score.

3 Experimental set-up

The IBM1 probabilities necessary for the IBM1 scores are learnt using the WMT 2010 News Commentary Spanish-English, French-English and German-English parallel texts. The language models are trained on the corresponding target parts of this corpus using the SRI language model tool (Stolcke, 2002). The POS tags for all languages were produced using the TreeTagger[1], and the morphemes are obtained using the Morfessor tool (Creutz and Lagus, 2005). The tool is corpus-based and language-independent: it takes a text as input and produces a segmentation of the word forms observed in the text. The obtained results are not strictly

[1] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

linguistic, however they often resemble a linguistic morpheme segmentation. Once a morpheme segmentation has been learnt from some text, it can be used for segmenting new texts. In our experiments, the splitting are learnt from the training corpus used for the IBM1 lexicon probabilities. The obtained segmentation is then used for splitting the corresponding source texts and hypotheses. Detailed corpus statistics are shown in Table 1.

Using the obtained probabilities, the scores described in Section 2 are calculated for the provided annotated data: the English-Spanish data from WMT 2008 consisting of four translation outputs produced by four different systems (Specia et al., 2010), the French-English and English-Spanish data from WMT 2010 (Specia, 2011), as well as for an additional WMT 2011 German-English and English-German annotated data. The human quality scores for the first two data sets range from 1 to 4, and for the third data set from 1 to 3. The interpretation of human scores is:

1. requires complete retranslation (*bad*)
2. post-editing quicker than retranslation ($edit^-$); this class was omitted for the third data set
3. little post-editing needed ($edit^+$)
4. fit for purpose (*good*)

As a first step, the arithmetic means and standard deviations are calculated for each feature and each class in order to see if the features are at all possible candidates for quality estimation, i.e. if the values for different classes are distinct.

After that, the main test is carried out: for each of the features, the Spearman correlation coefficient ρ with the human ranking are calculated for each document. In total, 9 correlation coefficients are obtained for each score – four Spanish outputs from the WMT 2008 task, one Spanish and one English output from the WMT 2010 as well as one English and two German outputs from the WMT 2011 task.

The obtained correlation results were then summarised into the following two values:

- *mean*
  a correlation coefficient averaged over all translation outputs;
- *rank>*
  percentage of translation outputs where the particular feature has better correlation than the other investigated features.

4 Results

4.1 Arithmetic means

The preliminary experiments consisted of comparing arithmetic means of scores for each feature and each class. The idea is: if the values are distinct enough, the feature is a potential candidate for quality estimation. In addition, standard deviations were calculated in order to estimate the overlapping.

For most translation outputs, all of our features have distinct arithmetic means for different classes and decent standard deviations, indicating that they are promising for further investigation. On all WMT 2011 outputs annotated with three classes, the distinction is rather clear, as well as for the majority of the four class outputs.

However, on some of the four class translation outputs, the values of the *bad* translation class were unexpected in the following two ways:

- the *bad* class overlaps with the $edit^-$ class;
- the *bad* class overlaps with the $edit^+$ class.

The first overlapping problem occured on two translation outputs of the 2011 set, and the second one on the both outputs of the 2010 set.

Examples for the PLM6 and P4IBM1$_{sh}$ features are shown in Table 2. First two rows present three class and four class outputs with separated arithmetic means, the first problem is shown in the third row, and the second (and more serious) problem is presented in the last row.

These overlaps have not been investigated further in the framework of this work, however this should be studied deeply (especially the second problem) in order to better understand the underlying phenomena and improve the features.

4.2 Spearman correlation coefficients

As mentioned in the previous section, Spearman rank correlation coefficients are calculated for each translation output and for each feature, and summarised into two values described in Section 3, i.e.

	Spanish	English	French	English	German	English
sentences	97122		83967		100222	
running words	2661344	2338495	2395141	2042085	2475359	2398780
vocabulary:						
words	69620	53527	56295	50082	107278	54270
morphemes	14178	13449	12004	12485	22211	13499
POS tags	69	44	33	44	54	44
POS-4grams	135166	121182	62177	114555	114314	123550

Table 1: Statistics of the corpora for training IBM1 lexicon models and language models.

feature	output / class	*ok*	*edit*$^+$	*edit*$^-$	*bad*
PLM6	de-en	13.5 / 7.3	23.7 / 13.6		33.0 / 19.7
	es-en4	10.9 / 5.0	20.7 / 8.7	34.6 / 16.4	49.0 / 23.7
	es-en3	18.5 / 11.0	30.2 / 15.6	**38.4 / 17.4**	**37.9 / 18.9**
	fr-en	15.2 / 8.8	**26.2 / 13.7**	34.5 / 18.4	**21.7 / 11.3**
P4IBM1$_{sh}$	de-en	50.5 / 38.4	109.7 / 75.6		161.8 / 108.3
	es-en4	37.9 / 25.0	88.7 / 48.7	165.8 / 89.0	241.5 / 127.4
	es-en3	77.0 / 56.7	139.8 / 82.5	**186.4 / 94.6**	**185.2 / 102.0**
	fr-en	53.5 / 44.3	**110.0 / 69.3**	151.8 / 90.9	**90.8 / 59.0**

Table 2: Arithmetic means with standard deviations of PLM6 and P4IBM1$_{sh}$ scores for four translation outputs: first two rows present decently separated classes, third row illustrates the overlap problem concerning the *bad* and the *edit*$^-$ class, the last row illustrates the overlap problem concerning the *bad* and the *edit*$^+$ class.

mean and *rank*>. The results are shown in Table 3. In can be seen that the best individual features are POS IBM1 scores followed by POS LM score.

The next step was to investigate combinations of the individual features. First, we calculated arithmetic mean of POS based features only, since they are more promising than the morpheme based ones, however we did not yield any improvements over the individual *mean* values. As a next step, we introduced weights to the features according to their mean correlations, i.e. we did not omit the morpheme features but put more weight on the POS based ones. Nevertheless, this also did not result in an improvement. Furthermore, we tried a simple arithmetic mean of all features, and this resulted in a better Spearman correlation coefficients.

Following all these observations, we decided to submit the arithmetic mean of all features to the WMT 2012 quality estimation task. Our submission consisted only of sentence ranking without scores, since we did not convert our scores to the interval [1,5]. Therefore we did not get any MAE or RMSE results, only DeltaAvg and Spearman correlation coefficients which were both 0.46. The highest scores in the shared task were 0.63, the lowest about 0.15, and for the "baseline" system which uses a set of well established features with an SVM classifier about 0.55.

5 Conclusions and outlook

The results presented in this article show that the IBM1 and the LM scores calculated on POS tags and morphemes have the potential to be used for the estimation of translation quality. These results are very preliminary, offering many directions for future work. The most important points are to use a classifier, as well as to combine the proposed features with already established features. Furthermore, the *bad* class overlapping problem described in Section 4.1 should be further investigated and understood.

Acknowledgments

This work has been partly developed within the TARAXÜ project financed by TSB Technologies-

mean		*rank>*	
0.449	P4IBM1$_{sh}$	70.4	P4IBM1$_{sh}$
0.445	P4IBM1$_{hs}$	68.5	P4IBM1$_{hs}$
0.444	PLM6	61.1	PLM6
0.430	MLM6	27.7	MLM6
0.426	MIBM1$_{sh}$	20.3	MIBM1$_{sh}$
0.420	MIBM1$_{hs}$	9.2	MIBM1$_{hs}$
0.450	arithmetic mean	**83.3**	arithmetic mean

Table 3: Features sorted by average correlation (column 1) and *rank>* value (column 2). The most promising score is the arithmetic mean of all individual features. The most promising individual features are POS-4gram IBM1 scores followed by POS-6gram language model score.

tiftung Berlin – Zukunftsfonds Berlin, co-financed by the European Union – European fund for regional development.

References

John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. 2003. Confidence estimation for machine translation. Final report, JHU/CLSP Summer Workshop.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311, June.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Omar Zaidan. 2011. Findings of the 2011 Workshop on Statistical Machine Translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation (WMT 2011)*, pages 22–64, Edinburgh, Scotland, July.

Mathias Creutz and Krista Lagus. 2005. Unsupervised morpheme segmentation and morphology induction from text corpora using morfessor 1.0. Technical Report Report A81, Computer and Information Science, Helsinki University of Technology, Helsinki, Finland, March.

Franz Josef Och, Daniel Gildea, Sanjeev Khudanpur, Anoop Sarkar, Kenji Yamada, Alex Fraser, Shankar Kumar, Libin Shen, David Smith, Katherine Eng, Viren Jain, Zhen Jin, and Dragomir Radev. 2003. Syntax for statistical machine translation. Technical report, Johns Hopkins University 2003 Summer Workshop on Language Engineering, Center for Language and Speech Processing, Baltimore, MD, USA, August.

Maja Popović, David Vilar Torres, Eleftherios Avramidis, and Aljoscha Burchardt. 2011. Evaluation without references: IBM1 scores as evaluation metrics. In *Proceedings of the Sixth Workshop on Statistical Machine Translation (WMT 2011)*, pages 99–103, Edinburgh, Scotland, July.

Maja Popović. 2011. Morphemes and POS tags for n-gram based evaluation metrics. In *Proceedings of the Sixth Workshop on Statistical Machine Translation (WMT 2011)*, pages 104–107, Edinburgh, Scotland, July.

Lucia Specia, Marco Turchi, Zhuoran Wang, John Shawe-Taylor, and Craig Saunders. 2009. Improving the confidence of machine translation quality estimates. In *Machine Translation Summit XII*, Ottawa, Canada.

Lucia Specia, Nicola Cancedda, and Marc Dymetman. 2010. A Dataset for Assessing Machine Translation Evaluation Metrics. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'2010)*, pages 3375–3378, Valletta, Malta, May.

Lucia Specia. 2011. Exploiting Objective Annotations for Measuring Translation Post-editing Effort. In *Proceedings of the 15th Annual Conference of the European Association for Machine Translation (EAMT 11)*, pages 73–80, Leuven, Belgium, May.

Drahomíra "Johanka" Spoustová, Jan Hajič, Jan Raab, and Miroslav Spousta. 2009. Semi-Supervised Training for the Averaged Perceptron POS Tagger. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 763–771, Athens, Greece, March.

Andreas Stolcke. 2002. SRILM – an extensible language modeling toolkit. In *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP 02)*, volume 2, pages 901–904, Denver, CO, September.