

Bag-of-Words Lexical Choice using Large Scale Classifiers

Patrick Haffner Srinivas Bangalore Stephan Kanthak
AT&T Labs - Research
180 Park Ave, Florham Park, NJ 07932
{haffner,srini,skanthak}@research.att.com

October 27, 2006

A major limitation in the application of Machine Learning to Machine Translation is that one needs to scale highly structured models to very large datasets, which is not possible with current optimization techniques. While current research applies structured methods to smaller datasets or subproblems, we take here an alternate path. In a first lexical choice step, we explore highly scalable unstructured classification methods that produce bags of words (BOW) rather than sentences. This would be enough for Cross-Language Information Retrieval. Full translation requires a second lexical ordering step that produces an “optimal” sequencing for the words in the bag. We demonstrate improvements in both translation accuracy (BLEU score) and word retrieval accuracy (F-measure) in various tasks

The goal of lexical choice is, given a source sentence, to estimate the probability that we find a given word in the target sentence. This is why, instead of producing a target sentence, what we initially obtain is a target BOW. Each word in the target vocabulary is detected independently, so we have here a very simple use of binary static classifiers. Training sentence pairs are considered as positive examples when the word appears in the target, and negative otherwise. Thus, the number number of training examples equals the number of sentence pair. The classifier is trained with n -gram features from the source sentence. Unlike most other statistical or discriminant MT approaches, there is no need to align the training sentences at the word level, using alignment algorithms such as GIZA++, which have been found to perform poorly for languages with radically different word order (e.g. English-Japanese).

During decoding, the BOW consists of the words with conditional probability greater than an adaptive threshold. In order to reconstruct the proper order of words in the target sentence we consider all permutations of words in the BOW and weight them by a target language model. This global reordering procedure uses a permutation automaton that grows exponentially with the sentence length, so limiting heuristics [3] are needed. To allow for length adjustments of target sentences, optional deletions are added in the final step of permutation decoding.

Our experiments compare two approaches. The baseline approach [1] builds a trigram model on this bilanguage corpus and represent the resulting model as a stochastic finite-state-transducer (SFST) which maps source language tokens to target language tokens. For decoding, we compose the source language test sentence with the SFST and search for the highest probability translation from the result of the composition. The BOW approach uses L1-regularized Maximum Entropy binary classifiers, which were shown [2] to combine state-of-the-art performance and superior scaling properties (when compared to SVM in particular).

	Dev 2005	Dev 2006	
	Text	Text	ASR 1-best
SFST	51.8	19.5	16.5
BOWMaxent	59.9	19.3	16.6

Table 1: Results (mBLEU) scores for the three different models on the transcriptions for development set 2005 and 2006 and ASR 1-best for development set 2006.

The IWSLT06 Chinese-English task is medium-size, with 46,311 training sentences. Results on the development sets from 2005 and 2006 are reported in Table 1. To assess how our lexical choice approach would do in a large scale retrieval task, we used a 1 million sentence-pair partition of the UN Arabic-English corpus as the training corpus and a different 994 sentence-pair partition as the test corpus. The Maxent-based model has 252,571 input features (unigrams, bigrams and trigrams of morphemes in the source Arabic sentences) and 53,005 classifiers (English words occurring more than 8 times). The retrieval accuracy of the decoded bag of target words is measured in terms of the F – *measure* (combination of recall and precision) against a reference bag of target words constructed from the target test sentence. While the baseline SFST approach yields a 64.6 F-measure, the BOWMaxent F-measure increases to 69.5. We do not have full translation results for this UN task yet.

The bag-of-words approach is very promising because it performs reasonably well despite considerable and easy to identify losses of information between the source and the target. The first and most obvious loss is about word position. The only information we use right now to restore the target word position is the global language model, the information about the position in the source sentence is completely lost. We are currently working toward incorporating syntactic information on the target words so as to be able to recover some of the position information lost in the classification process.

References

- [1] S. Bangalore and G. Riccardi. A Finite-State Approach to Machine Translation. In *NAACL, Pittsburgh*, 2001.
- [2] Patrick Haffner. Scaling Large Margin Classifiers for Spoken Language Understanding. In *Speech Communication*, 2005.
- [3] S. Kanthak, D. Vilar, E. Matusov, R. Zens, and H. Ney. Novel reordering approaches in phrase-based statistical machine translation. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 167–174, Ann Arbor, Michigan, 2005.