Multi-task Learning for Word Alignment and Dependency Parsing

Shujie Liu

Harbin Institute of Technology
No. 92, West Da-Zhi Street, Nangang District
Harbin City, China
shujieliu@mtlab.hit.edu.cn

Abstract. Word alignment and parsing are two important components for syntax based machine translation. The inconsistent models for alignment and parsing caused problems during translation pair extraction. In this paper, we do word alignment and dependency parsing in a multi-task learning framework, in which word alignment and dependency parsing are consistent and assisted with each other. Our experiments show significant improvement not only for both word alignment and dependency parsing, but also the final translation performance.

Keywords: Multi-task learning, word alignment, dependency parsing, machine translation.

1 Introduction

Since noisy channel model is introduced by Brown et al. (1994), statistical methods are widely-used in machine translation. Och (2003) proposed a more generalized log-linear model with minimum error rate training (MERT) which has provided substantial improvements over the original noisy channel model. This method uses a discriminative framework to integrate several sub-models: translation model, language model, distortion model, and so on. The sub-models can be generative or discriminative, such as a generative translation model and a discriminative maximum entropy distortion model. The sub-models are usually trained separately and then integrated into a log-linear framework with weights tuned using MERT.

Among the sub-models, translation model plays a key role which measures the faithfulness of a candidate as the translation of a source language sentence. Translation model is usually a translation table which contains translation pairs with translation probabilities. Most systems discover translation pairs via phrase extraction and maximum likelihood estimation (MLE). Before phrase extraction, the sentence pairs in the training data must be word aligned.

There are two kinds of approaches for word alignment. One is generative models, such as Giza++ using IBM Models and HMM model, and generative ITG models for word alignment. The generative models usually use EM method to train the model parameters. Another one is discriminative model with kinds of features, such as log-linear models used in [1], CRF models used in [2], and discriminative ITG models

H. Deng et al. (Eds.): AICI 2011, Part III, LNAI 7004, pp. 151–158, 2011.

used in [3] and [4]. One disadvantage of generative models is that it is not very easy to integrate kinds of features which are not independent with each other, while it is very simple and easy for discriminative models. Another reason of popularity of discriminative model is the discriminative model can optimize the parameters with the final evaluation metrics oriented.

Dependency grammar (DG) is a class of syntactic theories developed by Lucien Tesnière. It is distinct from phrase structure grammars, as it lacks phrasal nodes. Structure is determined by the relation between a word (a head) and its dependents. Dependency parsing is a task trying to find the dependency relations between words in a sentence.

One modern approach to building dependency parsers, called data-driven dependency parsing, is to learn good and bad parsing decisions solely from labeled data, without the intervention of an underlying grammar.

Since word alignment focuses the relations of the words of sentence pairs, and dependency parsing focuses the relations of the words in one sentence, there are many works studying the relations between word alignment and dependency parsing. Some work use the word alignment result to generate dependency tree of the target sentence given the dependency tree of the source sentence using projection methods, and also dependency information can be used to generate better word alignment result.

In this paper, word alignment and dependency parsing are integrated in a multi-task learning framework, in which word alignment and dependency parsing can be consistent and assisted with each other by introducing common feature for them. By integrating three tasks (word alignment, source sentence dependency parsing, target sentence dependency parsing), we not only achieved better word alignment result, better dependency result for source and target sentences, and also, the final translation performance is improved significantly.

In the following of this paper, we will introduce related works in section 2, followed by the introduction of multi-task learning in section 3. Our method will be shown in section 4 in detail, and the experiments are conducted in section 5. The final conclusion and future work will be discussed in section 6.

2 Related Work

Reference [6] proposed a projection method to get the dependency parsing result for the target language sentence using the dependency parsing result of the source language sentence and the word alignment result of this sentence pair. With a post-projection transformation, the f-score for the target language sentences can achieve comparable result with the result of a clean target language parser.

Reference [7] presented an empirical study that quantifies the degree to which syntactic dependencies are preserved when parses are projected directly from English to Chinese. Their results show that the quality of the projected Chinese parses can achieve F-score of 76% with a small set of principled, elementary linguistic transformations.

Reference [5] proposed a word alignment procedure based on a syntactic dependency analysis of French/English parallel corpora, which is called “alignment by syntactic propagation”: Both corpora are processed deeply with a dependency

parser, and then starting with an anchor word pair which has high confident translation probability, the alignment link is propagated to the syntactically connected words. Based on their experiments, this approach can achieve a precision of 94.3% and 93.1% with a recall of 58% and 56%, respectively for each corpus.

Reference [8] presented a new statistical model for computing the probability of an alignment given a sentence pair using dependency cohesion constraint features. The added dependency cohesion constraint features can achieve AER reduction of 1.8 points. Reference [9] introduced soft syntactic constraints into a discriminative ITG word alignment framework trained with SVM, and produced a 22% relative reduction in error rate with respect to a strong flat-string model.

The most similar work with ours is reference [10], in which they jointly parsed a bitext, and got improved parse quality on both sides. In a maximum entropy bitext paring model, they defined a distribution over source trees, target trees, and node-to-node alignments between them using monolingual parse scores and various measures of syntactic divergence. The resulting bitex parser outperforms state-of-the-art monolingual parser baseline by 2.5 F1 at predicting English side trees and 1.8 F1 at predicting Chinese side trees, and also these improved trees yielded significant improvement in final machine translation.

Compared with reference [10], our work is different with them in following ways: one is that our work is a combination process for parsing and word alignment, not started from scratch; the other one is our syntactic parser trees for source and target sentences are dependency trees instead of constituent trees in reference [10]; the third one is that our features used are different with those in reference [10].

3 Multi-task Learning

As in [11, 12], multi-task learning (MTL) is an inductive transfer mechanism whose principle goal is to improve generalization performance. MTL improves generalization by leveraging the domain-specific information contained in the training signals of related tasks. It does this by training tasks in parallel while using a shared representation.

The performance improvement from MTL is due to the extra information in the training signals of related tasks. The traditional learning process tries to learn each task from scratch, while multi-task learning tries to share the common knowledge among the tasks. MTL prefers distributions that other tasks prefer, and MTL prefers not to use distributions other task prefers not to use. Consider two tasks T_1 and T_2 sharing common knowledge, optimization process is biased towards representations in the intersection of what would be learned for T_1 and T_2.

4 Our Method

4.1 Model

Log-linear models are used to do word alignment and dependency parsing for sentence pairs. The log-linear mode for word alignment is shown in equation (1), where (e, f) is the sentence pair to be aligned and a is the word alignment candidate, and $\hat{a}$ is the

best candidate given by the log-linear model. f is the feature vector used in log-linear model for predicting the word alignment for the sentence pair (e, f), and λ is the weight vector for feature vector f:

$$\hat{a} = \arg\max_{a}(\sum_{i=1}^{n} \lambda_i f_i(a, e, f)) \tag{1}$$

And the log-linear model for dependency parsing is shown in equation (2), where e is the sentence to be parsed, d is the dependency parsing tree candidates, and $\hat{d}$ is the best parsing candidate given by equation (2):

$$\hat{d} = \arg\max_{d}(\sum_{i=1}^{n} \lambda_i f_i(d, e)) \tag{2}$$

With the multi-task learning framework, we integrate the three tasks (word alignment, source dependency parsing and target dependency parsing) into one log-linear model, which is shown in equation (3):

$$(\hat{a}, \hat{d}_s, \hat{d}_t) = \arg\max_{(a,ds,dt)}(\sum_{i=1}^{n} \lambda_i f_i(a, ds, dt, e, f) \tag{3}$$

where (ds, dt) are the dependency parsing trees for (e, f), and the features used in word alignment and source/target dependency trees are merged in to the new feature vector f with the feature weight λ.

Instead of starting from scratch to generate the word alignment matrix and dependency trees, here, we use the multi-task framework to do a combination task of results produced by other tools. For alignment, the alignment candidates for combination are generated by Giza++, an implementation of HMM alignment tool, and an implementation of a discriminative alignment in reference [13], and the dependency candidates for source and target sentences are generated from Berkeley parser and MST parser.

For the used word alignment tools, Giza++ is an implementation of IBM Models including IBM Model 1-5, the HMM alignment tool is an re-implement of HMM alignment, and the implementation of the discriminative alignment tool uses log-linear model and beam search to search a local optimized alignment result. The Berkeley parser is an un-lexicalized parser with hierarchically state-split PCFGs, with a coarse-to-fine method in which a grammar's own hierarchical projections are used to incremental pruning, and the MST parser is a non-projective dependency parser that searches for maximum spanning trees over directed graphs. Models of dependency structure are based on large-margin discriminative training methods.

With the output of initial alignment and dependency parsers, we use the intersection of the results as the start point link set of word alignment and dependency links, and then we perform a beam search process to add new links until the score of equation (3) is not improved. The added links must be links in one of the results for combining.

The used training method is minimum error rate training (MERT) as used in [15]. The objective function of the training target is defined as combination of F-score for word alignment and F-score for dependency parsing of source and target sentences.

4.2 Features for Word Alignment

- Translation probability
  The translation probability for a word pair (e_w, f_w) (e_w is a English word and f_w is Chinese word) is calculated using maximum-likelihood estimation (MLE), with the following equation:

$$P(e_w, f_w) = \frac{Count(e_w, f_w)}{\sum_{e_w'} Count(e_w', f_w)} \tag{4}$$

  where e_w' is any English word, $Count(e_w, f_w)$ is the count of times e_w is aligned to f_w. This count is calculated from the result of Giza++.
- Fertility probability
  The fertility probability depict the probability of a source word (English word) generates a number of Chinese words, as used in [14]. The fertility probability we used here are also calculated using MLE:

$$P(e_w, n) = \frac{Count(e_w, n)}{\sum_{m} Count(e_w, m)} \tag{5}$$

  where *n* and *m* are the number of fertility, and $Count(e_w, n)$ is the count of English word e_w are aligned to n Chinese words.
- Distortion probability
  The distortion probability is calculated in the same way as in [14].

4.3 Features for Dependent Parsing

- Dependency probability
  The dependency probability for a word pair is calculated using maximum-likelihood estimation (MLE), as similar as translation probability for word alignment.
- Root probability
  The root probability is the probability of a word as a root of a dependency tree, also calculated using MLE.
- Leaf probability
  The leaf probability is the probability of a word as a leaf of a dependency tree, which means this word can't be head of any words in the sentence, also calculated using MLE.

4.4 Features for Both Alignment and Parsing

- Dependency propagation
  Supposing there are two words A,B in English sentence which have a X dependency relation, and there are two words C,D in Chinese sentence which also have a X dependency relation, and A and C have a alignment link, if there's a alignment link between B and D, we call it a good dependency propagation. The dependency propagation feature is the count of good dependency propagations.
- Functional word links' suggestion
  For example in Figure 1, there is a Chinese functional "笔", and it's very hard to align it to the English word "that". But as we know there's dependency link between "那" and "笔", and this can be used to suggest a link of "that" and "笔". The functional word links' suggestion feature is the count of such phenomenon.

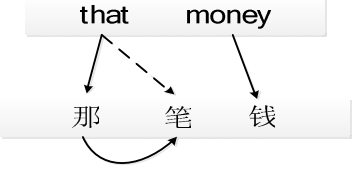

Fig. 1. An example of functional word links' suggestion

5 Experiments

5.1 Experiment Setting

The dependency tree for sentence pairs are LDC2007T02, we extracted 1000 sentence pairs and annotated with word alignment result. The first 500 sentence pairs are used as training data, and the left as the testing data. The original tree bank is constituent syntactic trees, so we use Penn2Malt[1] to convert the constituent tree to dependency tree.

5.2 Performance for Word Alignment and Dependency Parsing

The first experiment is to evaluate the word alignment performance and dependency parsing performance. The word alignment results are shown in Table 1.

From Table 1, we can find that the multi-task learning combination can improve the word alignment performance about 2 points, not only for precision, but also for recall.

The dependency parsing result are shown in Table 2.

From Table 2, the improvement of our approach is about 3 points better compared with the two baseline systems.

[1] http:// w3.msi.vxu.se/~nivre/research/Penn2Malt.html

Table 1. Word alignment performance using MTL. "Discrim" is the word alignment performance of the implementation of discriminative word alignment model as said in section Ⅳ.A. MTL is our approach with multi-task learning framework.

	Precision	Recall	F-Score
Giza++	0.84	0.82	0.83
HMM	0.79	0.84	0.82
Discrim	0.87	0.82	0.85
MTL	0.89	0.84	0.87

Table 2. Dependency parsing results using MTL combination. "Berkeley" stands for the dependency parsing result from Berkeley parser, and MST is the output of MST parser, while MTL is the dependency parsing result using MTL combination.

	Precision	Recall	F-Score
Berkeley	0.78	0.77	0.77
MST	0.79	0.80	0.89
MTL	0.81	0.83	0.82

5.3 Performance for Machine Translation

We also conduct an end-to-end evaluation of the alignment results with machine translation performance. The bilingual training dataset is the NIST training set excluding the Hong Kong Law and Hong Kong Hansard, and our 5-gram language model is trained from the Xinhua section of the Gigaword corpus. The NIST'03 test set is used as our development corpus and the NIST'05 and NIST'08 test sets are our test sets. We use our implementation of hierarchical phrase-based SMT (Chiang, 2007), with standard features. The SMT performance is shown in Table 3. From Table 3, we can find our method not only can improve the word alignment results and dependency parsing performance for source and target sentences, but also can improve the final machine translation performance on Nist'05 and Nist'06 significantly.

Table 3. Machine translation performance using the word alignment generated by MTL

	NIST'05	NIST'08
HMM	36.91	26.86
Giza	37.70	27.33
Discrim	37.51	27.42
MTL	38.32	27.95

6 Conclusion

In this paper, we do word alignment and dependency parsing in a multi-task learning framework, in which word alignment and dependency parsing are consistent and assisted with each other. Instead of starting from scratch to generate the alignment and dependency trees, we use the initial results from other tools to do a combination.

Our experiments show significant improvement not only for both word alignment and dependency parsing, but also the final translation performance.

References

1. Moore, R.C.: Sneddon, A discriminative framework for bilingual word alignment. In: HLT/EMNLP (2005)
2. Blunsom, P., Cohn, T.: Discriminative word alignment with conditional random fields. In: ACL (2006)
3. Cherry, C., Lin, D.: Soft Syntactic Constraints for Word alignment through discrinative training. In: ACL/COLING (2006)
4. Liu, S., Li, C.-H., Zhou, M.: Discriminative pruning for discriminative ITG alignment. In: ACL (2010)
5. Ozdowska, S.: Identifying correspondences between words: an approach based on a bilingual syntactic analysis of French/English parallel corpora. In: MLR 2004 Proceedings of the Workshop on Multilingual Linguistic Ressources (2004)
6. Hwa, R., Resnik, P., Weinberg, A.: Breaking the Resource Bottleneck for Multilingual Parsing. Technical report (2002)
7. Hwa, R., Resnik, P., Weinberg, A., Kolak, O.: Evaluating translational correspondence using Annotation Projection. In: ACL (2002)
8. Cherry, C., Lin, D.: A probability model to improve word aligment. In: ACL (2003)
9. Cherry, C., Lin, D.: Soft syntactic constraints for word alignment through discriminative training. In: COLING/ACL (2006)
10. Burkett, D., Klein, D.: Two languages are better than one (for syntactic parsing). In: EMNLP (2008)
11. Caruana, R.: Multitask Learning. In: Machine Learning (1997)
12. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering
13. Liu, Y., Liu, Q., Lin, S.: Log-linear models for word alignment. In: ACL (2005)
14. Brown, P.F., Della Pietra, S.A., Della Pietra, V.J., Mercer, R.L.: The mathematics of statistical machine translation: parameter estimation. In: ACL (2003)
15. Fraser, A., Marcu, D.: Semi-supervised training for statistical word alignment. In: ACL (2006)