Word Alignment with Synonym Regularization

Hiroyuki Shindo, Akinori Fujino, and **Masaaki Nagata**
NTT Communication Science Laboratories, NTT Corp.
2-4 Hikaridai Seika-cho Soraku-gun Kyoto 619-0237 Japan
{shindo,a.fujino}@cslab.kecl.ntt.co.jp
nagata.masaaki@lab.ntt.co.jp

Abstract

We present a novel framework for word alignment that incorporates synonym knowledge collected from monolingual linguistic resources in a bilingual probabilistic model. Synonym information is helpful for word alignment because we can expect a synonym to correspond to the same word in a different language. We design a generative model for word alignment that uses synonym information as a regularization term. The experimental results show that our proposed method significantly improves word alignment quality.

1 Introduction

Word alignment is an essential step in most phrase and syntax based statistical machine translation (SMT). It is an inference problem of word correspondences between different languages given parallel sentence pairs. Accurate word alignment can induce high quality phrase detection and translation probability, which leads to a significant improvement in SMT performance. Many word alignment approaches based on generative models have been proposed and they learn from bilingual sentences in an unsupervised manner (Vogel et al., 1996; Och and Ney, 2003; Fraser and Marcu, 2007).

One way to improve word alignment quality is to add linguistic knowledge derived from a monolingual corpus. This monolingual knowledge makes it easier to determine corresponding words correctly. For instance, functional words in one language tend to correspond to functional words in another language (Deng and Gao, 2007), and the syntactic dependency of words in each language can help the alignment process (Ma et al., 2008). It has been shown that such *grammatical* information works as a constraint in word alignment models and improves word alignment quality.

A large number of monolingual *lexical* semantic resources such as WordNet (Miller, 1995) have been constructed in more than fifty languages (Sagot and Fiser, 2008). They include word-level relations such as synonyms, hypernyms and hyponyms. Synonym information is particularly helpful for word alignment because we can expect a synonym to correspond to the same word in a different language. In this paper, we explore a method for using synonym information effectively to improve word alignment quality.

In general, synonym relations are defined in terms of word sense, not in terms of word form. In other words, synonym relations are usually context or domain dependent. For instance, 'head' and 'chief' are synonyms in contexts referring to working environment, while 'head' and 'forefront' are synonyms in contexts referring to physical positions. It is difficult, however, to imagine a context where 'chief' and 'forefront' are synonyms. Therefore, it is easy to imagine that simply replacing all occurrences of 'chief' and 'forefront' with 'head' do sometimes harm with word alignment accuracy, and we have to model either the context or senses of words.

We propose a novel method that incorporates synonyms from monolingual resources in a bilingual word alignment model. We formulate a synonym pair generative model with a topic variable and use this model as a regularization term with a bilingual word alignment model. The topic variable in our synonym model is helpful for disambiguating the meanings of synonyms. We extend HM-BiTAM, which is a HMM-based word alignment model with a latent topic, with a novel synonym pair generative model. We applied the proposed method to an English-French word alignment task and successfully improved the word

Proceedings of the ACL 2010 Conference Short Papers, pages 137–141,
Uppsala, Sweden, 11-16 July 2010. ©2010 Association for Computational Linguistics

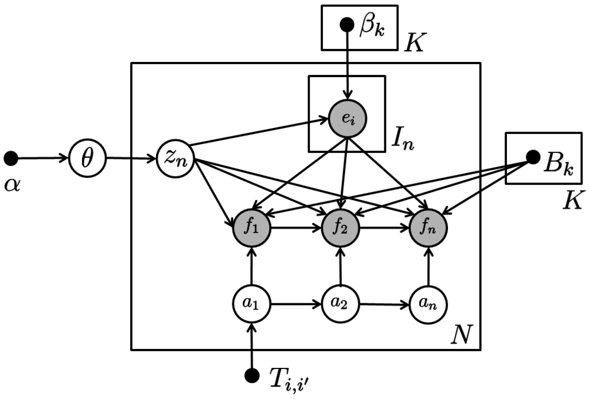

Figure 1: Graphical model of HM-BiTAM

alignment quality.

2 Bilingual Word Alignment Model

In this section, we review a conventional generative word alignment model, HM-BiTAM (Zhao and Xing, 2008).

HM-BiTAM is a bilingual generative model with topic z, alignment a and topic weight vector θ as latent variables. Topic variables such as ‘science’ or ‘economy’ assigned to individual sentences help to disambiguate the meanings of words. HM-BiTAM assumes that the nth bilingual sentence pair, (E_n, F_n), is generated under a given latent topic $z_n \in \{1, \ldots, k, \ldots, K\}$, where K is the number of latent topics. Let N be the number of sentence pairs, and I_n and J_n be the lengths of E_n and F_n, respectively. In this framework, all of the bilingual sentence pairs $\{E, F\} = \{(E_n, F_n)\}_{n=1}^{N}$ are generated as follows.

1. $\theta \sim Dirichlet\,(\alpha)$: sample topic-weight vector
2. For each sentence pair (E_n, F_n)
   (a) $z_n \sim Multinomial\,(\theta)$: sample the topic
   (b) $e_{n,i:I_n}\,|z_n \sim p\,(E_n\,|z_n;\beta)$: sample English words from a monolingual unigram model given topic z_n
   (c) For each position $j_n = 1, \ldots, J_n$
      i. $a_{j_n} \sim p\,(a_{j_n}\,|a_{j_n-1};T)$: sample an alignment link a_{j_n} from a first order Markov process
      ii. $f_{j_n} \sim p\,(f_{j_n}\,|E_n, a_{j_n}, z_n;B)$: sample a target word f_{j_n} given an aligned source word and topic

where alignment $a_{j_n} = i$ denotes source word e_i and target word f_{j_n} are aligned. α is a parameter over the topic weight vector θ, $\beta = \{\beta_{k,e}\}$ is the source word probability given the kth topic: $p\,(e\,|z = k)$. $B = \{B_{f,e,k}\}$ represents the word translation probability from e to f under the kth topic: $p\,(f\,|e, z = k)$. $T = \{T_{i,i'}\}$ is a state transition probability of a first order Markov process. Fig. 1 shows a graphical model of HM-BiTAM.

The total likelihood of bilingual sentence pairs $\{E, F\}$ can be obtained by marginalizing out latent variables z, a and θ,

$$p\,(F, E; \Psi) = \sum_{z}\sum_{a}\int p\,(F, E, z, a, \theta; \Psi)\,d\theta, \qquad (1)$$

where $\Psi = \{\alpha, \beta, T, B\}$ is a parameter set. In this model, we can infer word alignment a by maximizing the likelihood above.

3 Proposed Method

3.1 Synonym Pair Generative Model

We design a generative model for synonym pairs $\{f, f'\}$ in language F, which assumes that the synonyms are collected from monolingual linguistic resources. We assume that each synonym pair (f, f') is generated independently given the same ‘sense’ s. Under this assumption, the probability of synonym pair (f, f') can be formulated as,

$$p\,(f, f') \propto \sum_{s} p\,(f\,|s)\,p\,(f'\,|s)\,p\,(s)\,. \qquad (2)$$

We define a pair (e, k) as a representation of the sense s, where e and k are a word in a different language E and a latent topic, respectively. It has been shown that a word e in a different language is an appropriate representation of s in synonym modeling (Bannard and Callison-Burch, 2005). We assume that adding a latent topic k for the sense is very useful for disambiguating word meaning, and thus that (e, k) gives us a good approximation of s. Under this assumption, the synonym pair generative model can be defined as follows.

$$\begin{aligned} &p\left(\{f, f'\}\,; \widetilde{\Psi}\right) \\ &\propto \prod_{(f,f')}\sum_{e,k} p(f|e,k;\widetilde{\Psi})p(f'|e,k;\widetilde{\Psi})p(e,k;\widetilde{\Psi}), \end{aligned} \qquad (3)$$

where $\widetilde{\Psi}$ is the parameter set of our model.

3.2 Word Alignment with Synonym Regularization

In this section, we extend the bilingual generative model (HM-BiTAM) with our synonym pair model. Our expectation is that synonym pairs

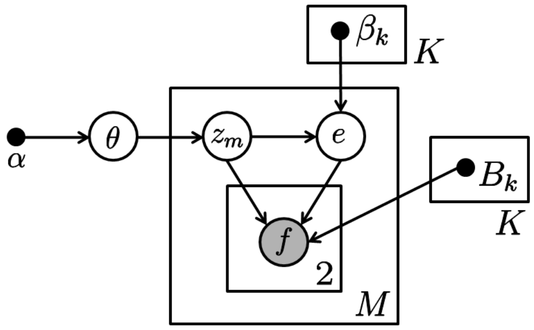

Figure 2: Graphical model of synonym pair generative process

correspond to the same word in a different language, thus they make it easy to infer accurate word alignment. HM-BiTAM and the synonym model share parameters in order to incorporate monolingual synonym information into the bilingual word alignment model. This can be achieved via reparameterizing $\widetilde{\Psi}$ in eq. 3 as,

$$p\left(f\left|e,k;\widetilde{\Psi}\right.\right) \equiv p\left(f\left|e,k;B\right.\right), \quad (4)$$

$$p\left(e,k;\widetilde{\Psi}\right) \equiv p\left(e\left|k;\beta\right.\right)p\left(k;\alpha\right). \quad (5)$$

Overall, we re-define the synonym pair model with the HM-BiTAM parameter set Ψ,

$$p(\{f,f'\};\Psi) \propto \frac{1}{\sum_{k'}\alpha_{k'}}\prod_{(f,f')}\sum_{k,e}\alpha_k\beta_{k,e}B_{f,e,k}B_{f',e,k}. \quad (6)$$

Fig. 2 shows a graphical model of the synonym pair generative process. We estimate the parameter values to maximize the likelihood of HM-BiTAM with respect to bilingual sentences and that of the synonym model with respect to synonym pairs collected from monolingual resources. Namely, the parameter estimate, $\hat{\Psi}$, is computed as

$$\hat{\Psi} = \arg\max_{\Psi}\left\{\log p(F,E;\Psi) + \zeta\log p(\{f,f'\};\Psi)\right\}, \quad (7)$$

where ζ is a regularization weight that should be set for training. We can expect that the second term of eq. 7 to constrain parameter set Ψ and avoid overfitting for the bilingual word alignment model. We resort to the variational EM approach (Bernardo et al., 2003) to infer $\hat{\Psi}$ following HM-BiTAM. We omit the parameter update equation due to lack of space.

4 Experiments

4.1 Experimental Setting

For an empirical evaluation of the proposed method, we used a bilingual parallel corpus of English-French Hansards (Mihalcea and Pedersen, 2003). The corpus consists of over 1 million sentence pairs, which include 447 manually word-aligned sentences. We selected 100 sentence pairs randomly from the manually word-aligned sentences as development data for tuning the regularization weight ζ, and used the 347 remaining sentence pairs as evaluation data. We also randomly selected 10k, 50k, and 100k sized sentence pairs from the corpus as additional training data. We ran the unsupervised training of our proposed word alignment model on the additional training data and the 347 sentence pairs of the evaluation data. Note that manual word alignment of the 347 sentence pairs was not used for the unsupervised training. After the unsupervised training, we evaluated the word alignment performance of our proposed method by comparing the manual word alignment of the 347 sentence pairs with the prediction provided by the trained model.

We collected English and French synonym pairs from WordNet 2.1 (Miller, 1995) and WOLF 0.1.4 (Sagot and Fiser, 2008), respectively. WOLF is a semantic resource constructed from the Princeton WordNet and various multilingual resources. We selected synonym pairs where both words were included in the bilingual training set.

We compared the word alignment performance of our model with that of GIZA++ 1.03 [1] (Vogel et al., 1996; Och and Ney, 2003), and HM-BiTAM (Zhao and Xing, 2008) implemented by us. GIZA++ is an implementation of IBM-model 4 and HMM, and HM-BiTAM corresponds to $\zeta = 0$ in eq. 7. We adopted $K = 3$ topics, following the setting in (Zhao and Xing, 2006).

We trained the word alignment in two directions: English to French, and French to English. The alignment results for both directions were refined with 'GROW' heuristics to yield high precision and high recall in accordance with previous work (Och and Ney, 2003; Zhao and Xing, 2006). We evaluated these results for precision, recall, F-measure and alignment error rate (AER), which are standard metrics for word alignment accuracy (Och and Ney, 2000).

[1] http://fjoch.com/GIZA++.html

10k		Precision	Recall	F-measure	AER
GIZA++	standard	0.856	0.718	0.781	0.207
	with SRH	0.874	0.720	0.789	0.198
HM-BiTAM	standard	0.869	0.788	0.826	0.169
	with SRH	0.884	0.790	0.834	0.160
Proposed		**0.941**	**0.808**	**0.870**	**0.123**

(a)

50k		Precision	Recall	F-measure	AER
GIZA++	standard	0.905	0.770	0.832	0.156
	with SRH	0.903	0.759	0.825	0.164
HM-BiTAM	standard	0.901	0.814	0.855	0.140
	with SRH	0.899	0.808	0.853	0.145
Proposed		**0.947**	**0.824**	**0.881**	**0.112**

(b)

100k		Precision	Recall	F-measure	AER
GIZA++	standard	0.925	0.791	0.853	0.136
	with SRH	**0.934**	0.803	0.864	0.126
HM-BiTAM	standard	0.898	0.851	0.874	0.124
	with SRH	0.909	0.860	0.879	0.114
Proposed		0.927	**0.862**	**0.893**	**0.103**

(c)

Table 1: Comparison of word alignment accuracy. The best results are indicated in bold type. The additional data set sizes are (a) 10k, (b) 50k, (c) 100k.

# vocabularies		10k	50k	100k
English	standard	8578	16924	22817
	with SRH	5435	7235	13978
French	standard	10791	21872	30294
	with SRH	9737	20077	27970

Table 2: The number of vocabularies in the 10k, 50k and 100k data sets.

4.2 Results and Discussion

Table 1 shows the word alignment accuracy of the three methods trained with 10k, 50k, and 100k additional sentence pairs. For all settings, our proposed method outperformed other conventional methods. This result shows that synonym information is effective for improving word alignment quality as we expected.

As mentioned in Sections 1 and 3.1, the main idea of our proposed method is to introduce *latent topics* for modeling synonym pairs, and then to utilize the synonym pair model for the regularization of word alignment models. We expect the latent topics to be useful for modeling polysemous words included in synonym pairs and to enable us to incorporate synonym information effectively into word alignment models. To confirm the effect of the synonym pair model with latent topics, we also tested GIZA++ and HM-BiTAM with what we call *Synonym Replacement Heuristics (SRH)*, where all of the synonym pairs in the bilingual training sentences were simply replaced with a representative word. For instance, the words 'sick' and 'ill' in the bilingual sentences were replaced with the word 'sick'. As shown in Table 2, the number of vocabularies in the English and French data sets decreased as a result of employing the SRH.

We show the performance of GIZA++ and HM-BiTAM with the SRH in the lines entitled "with SRH" in Table 1. The GIZA++ and HM-BiTAM with the SRH slightly outperformed the *standard* GIZA++ and HM-BiTAM for the 10k and 100k data sets, but underperformed with the 50k data set. We assume that the SRH mitigated the overfitting of these models into low-frequency word pairs in bilingual sentences, and then improved the word alignment performance. The SRH regards all of the different words coupled with the same word in the synonym pairs as synonyms. For instance, the words 'head', 'chief' and 'forefront' in the bilingual sentences are replaced with 'chief', since ('head', 'chief') and ('head', 'forefront') are synonyms. Obviously, ('chief', 'forefront') are not synonyms, which is detrimented to word alignment.

The proposed method consistently outperformed GIZA++ and HM-BiTAM with the SRH in 10k, 50k and 100k data sets in F-measure. The synonym pair model in our proposed method can automatically learn that ('head', 'chief') and ('head', 'forefront') are individual synonyms with different meanings by assigning these pairs to different topics. By sharing latent topics between the synonym pair model and the word alignment model, the synonym information incorporated in the synonym pair model is used directly for training word alignment model. The experimental results show that our proposed method was effective in improving the performance of the word alignment model by using synonym pairs including such *ambiguous* synonym words.

Finally, we discuss the data set size used for unsupervised training. As shown in Table 1, using a large number of additional sentence pairs improved the performance of all the models. In all our experimental settings, all the additional sentence pairs and the evaluation data were selected from the Hansards data set. These experimental results show that a larger number of sentence pairs was more effective in improving word alignment performance when the sentence pairs were collected from a *homogeneous* data source. However, in practice, it might be difficult to collect a large number of such homogeneous sentence pairs for a specific target domain and language pair. One direction for future work is to confirm the effect of the proposed method when training the word alignment model by using a large number of sentence pairs collected from various data sources including many topics for a specific language pair.

5 Conclusions and Future Work

We proposed a novel framework that incorporates synonyms from monolingual linguistic resources in a word alignment generative model. This approach utilizes both bilingual and monolingual synonym resources effectively for word alignment. Our proposed method uses a latent topic for bilingual sentences and monolingual synonym pairs, which is helpful in terms of word sense disambiguation. Our proposed method improved word alignment quality with both small and large data sets. Future work will involve examining the proposed method for different language pairs such as English-Chinese and English-Japanese and evaluating the impact of our proposed method on SMT performance. We will also apply our proposed method to a larger data sets of multiple domains since we can expect a further improvement in word alignment accuracy if we use more bilingual sentences and more monolingual knowledge.

References

C. Bannard and C. Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 597–604. Association for Computational Linguistics Morristown, NJ, USA.

J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West. 2003. The variational bayesian EM algorithm for incomplete data: with application to scoring graphical model structures. In *Bayesian Statistics 7: Proceedings of the 7th Valencia International Meeting, June 2-6, 2002*, page 453. Oxford University Press, USA.

Y. Deng and Y. Gao. 2007. Guiding statistical word alignment models with prior knowledge. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 1–8, Prague, Czech Republic, June. Association for Computational Linguistics.

A. Fraser and D. Marcu. 2007. Getting the structure right for word alignment: LEAF. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 51–60, Prague, Czech Republic, June. Association for Computational Linguistics.

Y. Ma, S. Ozdowska, Y. Sun, and A. Way. 2008. Improving word alignment using syntactic dependencies. In *Proceedings of the ACL-08: HLT Second Workshop on Syntax and Structure in Statistical Translation (SSST-2)*, pages 69–77, Columbus, Ohio, June. Association for Computational Linguistics.

R. Mihalcea and T. Pedersen. 2003. An evaluation exercise for word alignment. In *Proceedings of the HLT-NAACL 2003 Workshop on building and using parallel texts: data driven machine translation and beyond-Volume 3*, page 10. Association for Computational Linguistics.

G. A. Miller. 1995. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):41.

F. J. Och and H. Ney. 2000. Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 440–447. Association for Computational Linguistics Morristown, NJ, USA.

F. J. Och and H. Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

B. Sagot and D. Fiser. 2008. Building a free French wordnet from multilingual resources. In *Proceedings of Ontolex*.

S. Vogel, H. Ney, and C. Tillmann. 1996. HMM-based word alignment in statistical translation. In *Proceedings of the 16th Conference on Computational Linguistics-Volume 2*, pages 836–841. Association for Computational Linguistics Morristown, NJ, USA.

B. Zhao and E. P. Xing. 2006. BiTAM: Bilingual topic admixture models for word alignment. In *Proceedings of the COLING/ACL on Main Conference Poster Sessions*, page 976. Association for Computational Linguistics.

B. Zhao and E. P. Xing. 2008. HM-BiTAM: Bilingual topic exploration, word alignment, and translation. In *Advances in Neural Information Processing Systems 20*, pages 1689–1696, Cambridge, MA. MIT Press.