Meaning Representations in Statistical Word Alignment

Tsuyoshi Okita
School of Computing
Dublin City University
Glasnevin, Dublin 9, Ireland
tokita@computing.dcu.ie

1 Statistical Word Alignment Implemented by Graphical Model

Using an architecture of our statistical word alignment, this paper presents 1) how meaning representations is structured in our system which intends to be easily interpreted by a computer and still express rich / complex / conflicting knowledge, and 2) how we jointly infer semantics from several fragmental evidences of semantics (modalities).

As a testbed of statistical word aligner, we implemented the prototype of statistical word aligner by graphical models [2, 10]. The advantage of using graphical method resides in its extensibility compared to the traditional approach for statistical word alignment [3, 22, 14]. Although there are semi-supervised word aligner [6], we only talk about unsupervised word aligner [3, 22, 14]. The capabilities of this word aligner include that 1) it supports IBM / HMM models as well as tree-based Models [13], 2) it can extend easily to support MAP assignment-based decoder (Viterbi decoding [21] as well as posterior decoding [10]) in these models [16, 15], 3) it can be used for the input in the lattice-based decoding [1, 4] which are reinterpreted as the partial model selection, 4) it supports flexible on / off capability of random variables which has advantageous in the lemma-based alignment [5] and the morpheme avoided alignment, and 4) it can be used for the forced alignment [18]. This comes from the fact that the inference algorithms, such as sum-product and max-product algorithms, are not affected by the form of network structures. Note that the traditional statistical word alignment is built purely counting the frequency of words where syntax / semantics are not considered [11].

Let us e denote English word, f denote French word, and a denote alignment function [3, 8]. Let a_i denote the alignment function mapped from the i-th word in f into some word in e, f_T denote the (dependency) tree structured input in f, and e_T denote the (dependency) tree structured input in e. e, f are random variables and a is hidden variable.

1. (Independent model) EM-based word alignment is the basis of other variations.

$$\max \mathbb{E} p(f,a|e) = \frac{1}{Z}\sum_{a_1=0}^{l}\dots\sum_{a_m=0}^{l}\prod_{j=1}^{m} p(f_j|e_{a_j}) \quad \text{s.t.} \sum_{f} p(f|e) = 1 \tag{1}$$

2. (Chain model) HMM-based word alignment considers the first-order Markov dependency [22].

$$\max \mathbb{E} p(f,a|e) = \frac{1}{Z}\max_{a_1^m}\prod_{i=1}^{m} p(a_i|a_{i-1})p(f|e,a) \quad \text{s.t.} \sum_{f} p(f|e) = 1 \tag{2}$$

3. (Tree model) Tree-based word alignment consider the dependency structures in the source and the target language [13].

$$\max \mathbb{E} p(f,a|e) = \frac{1}{Z}\max_{a_1^m}\prod_{T\in\mathbb{T}} p(f_T|e_T,a)p(a|e_T)\prod_{T\in\mathbb{T}} p(e_T|f_T,a)p(a|f_T) \quad \text{s.t.} \sum_{T\in\mathbb{T}} p(f_T|e_T) = 1 \tag{3}$$

It is noted that the above formulation omits the description of MAP inference approach [10, 12]. In MAP inference, the random variables e, f are further partitioned into E (evidence) / Q (query) / H (hidden /nuisance), that is e_E, e_Q, f_E, and f_Q.

2 Meaning Representation

Instead of searching structured patterns from the beginning, we take the order of searching similar as the frequent graph mining [20]: first we search the frequent subgraphs, and then we combine them to find structured patterns. The feasibility of this approach in word alignment is that the subgraph objects that we listed in below, such as Multi Word Expressions (MWEs), pronoun, the coordinated Noun Phrases (NPs), semantic role, and so on, are reasonably easily detectable by current NLP technologies. Based on the detected subgraph objects, we write semantic action based on the type of subgraph objects. We explicitly write semantic actions when such subgraph objects appear.

Firstly, our mechanical interface of semantics /syntax / information structure is limited in the 'prior' in the MAP inference: we showed three types of word alignment models in the previous section. We use the prior which indicates the alignment links between e and f in sentence i [16].

Secondly, we define the meaning representation which supports the existence of alignment links with quantifying by probabilities: 1) alignment links $x_1, \cdots, x_m$ are (equally or with some probabilities) possible, 2) alignment links $x_1, \cdots, x_m$ is prohibited (or less likely), and 3) alignment link x is likely.

Thirdly, each of such semantic action will lead to the suggested word alignment links (whether it is 1-to-1, many-to-1, or many-to-many). Current version supports the following semantic actions: 1) MWEs invokes MWE-TM unit, 2) pronoun (and no subject) invokes referent-table unit, 3) IS-adverb invokes independent-IS-table unit, 4) the coordinated NPs invoke the coordination unit, 5) semantic role invokes the SR-table unit, 6) proper noun / transliteration / localization term / equation invokes the less-frequent-pair unit, 7) lexical semantics invokes the lexical-pair unit, 8) noise invokes the noise unit (noise is important for Japanese / Chinese), and so forth. Each of these is reasonably detected by current NLP technologies although this depends on the availability of such tools or corpora on the specific language. In our experiments, MWE detection is unsupervised learning, while POS-tagging, SRL, named-entity recognition are supervised learning. Note that currently the SR-table unit is only invokes semantic action when the semantic role in two languages are identical, and the noise unit simply mutates itself with its semantic action.

3 Inference

Each of these semantic action is converted into the suggested alignment links. Then, we combine such suggested alignment links by noisy-or to form the prior, which is supplied as prior knowledge to the MAP-based word aligner. Hence, the MAP inference follows the MAP assignment version of (1), (2) and (3) for given priors.

Note that the incorrect detection of subgraph object will lead to the incorrect invocation of the semantic action. Note also that despite that the alignment link should be natively binary value, we often set the prior not with binary value (0 or 1) but with probability.

4 Experiments

We evaluate the performance by BLEU [17] using EN-JP corpus [7] of 200k sentence pairs. We use the above word aligner, SRILM [19], and Moses [9]. Firstly, when we apply only a small amount of prior knowledge about alignment links, the effect of this is considerably small. However, once this surpasses some threshold, the effect becomes radically big. In this sense, one interest of ours is what is a threshold and how much (prior) knowledge we need to supply in order to achieve such threshold. In a sentence, when 50-60% of alignment links is a priori detected, the result reaches precision of 97%.

References

[1] Bertoldi, N., Zens, R., and Federico, M. Speech translation by confusion network decoding. *ICASSP* (2007), 1297–1300.

[2] Bishop, C. M. Pattern recognition and machine learning. *Springer* (2006).

[3] Brown, P. F., Pietra, V. J., A.D.Pietra, S., and Mercer, R. L. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics, Vol.19, Issue 2* (1993), 263–311.

[4] Dyer, C., Muresan, S., and Resnik, P. Generalizing word lattice translation. *In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL-08)* (2008), 1012–1020.

[5] Fishel, M., Bojar, O., Zeman, D., and Berka, J. Automatic translation error analysis. *In the proceedings of TSD 2011* (2011).

[6] Fraser, A., and Marcu, D. 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)* (2007), 51–60.

[7] Fujii, A., Utiyama, M., Yamamoto, M., Utsuro, T., Ehara, T., Echizen-ya, H., and Shimohata, S. Overview of the patent translation task at the NTCIR-8 workshop. *In Proceedings of the 8th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access* (2010), 293–302.

[8] Koehn, P. Statistical machine translation. *Cambridge University Press* (2010).

[9] Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. Moses: Open source toolkit for Statistical Machine Translation. *In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions* (2007), 177–180.

[10] Koller, D., and Friedman, N. Probabilistic graphical models: Principles and techniques. *MIT Press* (2009).

[11] Ma, Y., Lambert, P., and Way, A. Tuning syntactically enhanced word alignment for statistical machine translation. *In Proceedings of the 13th Annual Meeting of the European Association for Machine Translation(EAMT 2009)* (2009), 250–257.

[12] Murphy, K. P. Lecture slides at university of british columbia. *http://www.cs.ubc.ca/ murphyk/* (2011).

[13] Nakazawa, T., and Kurohashi, S. Statistical phrase alignment model using dependency relation probability. *In Proceedings of the third Workshop on Syntax and Structure in Statistical Translation (SSST-3)* (2009), 10–18.

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

[15] Okita, T., Graham, Y., and Way, A. Gap between theory and practice: Noise sensitive word alignment in machine translation. *In Proceedings of the Workshop on Applications of Pattern Analysis (WAPA2010). Cumberland Lodge, England.* (2010).

[16] Okita, T., Guerra, A. M., Graham, Y., and Way, A. Multi-Word Expression sensitive word alignment. *In Proceedings of the Fourth International Workshop On Cross Lingual Information Access (CLIA2010, collocated with COLING2010), Beijing, China.* (2010), 1–8.

[17] Papineni, K., Roukos, S., Ward, T., and Zhu, W. BLEU: A Method For Automatic Evaluation of Machine Translation. *In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-02)* (2002).

[18] Schwartz, L. Multi-source translation methods. In *Proc. AMTA* (October 2008).

[19] Stolcke, A. SRILM – An extensible language modeling toolkit. *In Proceedings of the International Conference on Spoken Language Processing* (2002), 901–904.

[20] Tsuda, K., and Kudo, T. Clustering graphs by weighted substructure mining. *In Proceedings of the 23rd International Conference on Machine Learning (ICML)* (2006), 953–960.

[21] Viterbi, A. J. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory 13*, 2 (1967), 260–269.

[22] Vogel, S., Ney, H., and Tillmann, C. HMM-Based word alignment in statistical translation. *In Proceedings of the 16th International Conference on Computational Linguistics (COLING 1996)* (1996), 836–841.