Minimum Error Rate Training for Bilingual News Alignment

Can Wang, Yang Liu, Maosong Sun
Department of Computer Science and Technology
State Key Lab on Intelligent Technology and Systems
National Lab for Information Science and Technology
Tsinghua University, Beijing 100084, China
acanthu@gmail.com, liuyang2011@tsinghua.edu.cn, sms@tsinghua.edu.cn

Abstract

News articles in different languages on the same event are invaluable for analyzing standpoints and viewpoints in different countries. The major challenge to identify such closely related bilingual news articles is how to take full advantage of various information sources such as length, translation equivalence and publishing date. Accordingly, we propose a discriminative model for bilingual news alignment, which is capable of incorporating arbitrary information sources as features. Chinese word segmentation, Part-of-speech tagging and Named Entity Recognition technologies are used to calculate the semantic similarities between words or text as feature values. The feature weights are optimized using the minimum error rate training algorithm to directly correlate training objective to evaluation metric. Experiments on Chinese-English data show that our method significantly outperforms two strong baseline systems by 12.7% and 2.5%, respectively.

1 Introduction

A hot event usually leads to many news articles written in different languages from multiple sources, which often exhibit diverse standpoints and viewpoints. While identifying, analyzing, and summarizing such diversity from multilingual news articles is of great value, how to find multilingual news articles on the same event turns out to be the first obstacle. Therefore, there is an urgent need for multilingual news alignment: given news articles in different languages, find out the correspondence.

As a special case of multilingual news alignment, bilingual news alignment has attracted attention from a number of authors. Yang et al. (1997) propose to use cross-lingual information retrieval techniques to identify the correspondence between multilingual texts. Steinberger et al. (2002) calculate the semantic similarity of news in different languages using a multilingual thesaurus. Vu et al. (2009) present a feature-based method to include useful information sources. However, these efforts suffer from two major drawbacks:

* The alignment model is not discriminatively trained. Information sources usually have different contributions to predicting alignment. It is important to recognize such differences to take full advantage of information sources.

* The optimization objective is not directly related to evaluation metrics. Most previous methods are heuristic or generative. Even for discriminative methods, the optimization objectives are usually maximum likelihood or maximum a posteri, which are not directly related to evaluation metrics. Therefore, maximizing likelihood, posterior or heuristic does not necessarily result in the maximum in terms of evaluation metric.

To alleviate these problems, we propose a linear model for bilingual news alignment. The model is capable of incorporating arbitrary information sources as features to predict the alignment of bilingual news articles. Each feature is associated with a weight to represent the degree of importance. We use the minimum error rate training (MERT) algorithm (Och, 2003) to optimize feature weights with respect to evaluation metrics directly. Experiments on Chinese-English data show that our method significantly outperforms two strong baseline systems by 12.7% and 2.5%, respectively.

2 Related Work

The core task of bilingual news alignment is how to evaluate the similarity between two news articles written in different languages. And the difficulty is how to overcome the semantic ambiguities and language barriers.

Similarity between monolingual documents is usually measured by metrics such as cosine similarity, Jaccard coefficient or Pearson correlation coefficient (Huang et al., 2008). But these lexical similarity methods cannot always identify the semantic similarity of texts like "I own a dog" and "I have an animal" (Mihalcea et al., 2006). Then researchers try to utilize corpus or knowledge like WordNet to design semantic metrics for any two words and then calculate the semantic similarity between texts.

Similarity between bilingual documents is more difficult. Yang et al. (1997) use cross-lingual information retrieval techniques such as example-based term translation method, generalized vector space model (GVSM), latent semantic indexing (LSI) method to identify the correspondence between bilingual texts. Leek et al. (1999) use machine translation tools to help cross-language topic tracking. Steinberger et al. (2002) represent European document contents using descriptor terms of a multilingual thesaurus EUROVOC and measure the semantic similarity based on the distance between the two documents' representations. Vu et al. (2009) use Discrete Fourier Transform (DFT) score of a word's frequency chain to measure time distribution similarity as a feature to evaluate the bilingual news similarity.

Because of the differences in data sets, the performances of cross-language information retrieval or bilingual document alignment are different, but they are always poor, and worse on Chinese-English test. Vu et al. (2009) achieve a precision of 31.5% in English-Chinese corpora, and 63.4% in English-Malay corpora on Top-1 retrieval test.

Previous methods, whether monolingual or bilingual, all try to represent documents into a unified semantic space, and then calculate the similarity. But they are either not for the news alignment or cannot make full use of the news articles' characteristics. We propose a linear model, and incorporate useful information sources as features such as bilingual vector space model cosine similarity, text graph similarity, named entity similarity, news publishing date interval. We use the minimum error rate training (MERT) algorithm to optimize feature weights with respect to evaluation metrics directly, and achieve a precision of 58.63%, which significantly outperforms two strong baseline systems.

The rest of this paper is organized as follows. In section 3, we give a formal description of our model and training method. Section 4 describes the features and the score functions we use. In Section 5, we evaluate our model in Chinese-English news alignment task. Section 6 points to a conclusion.

3 Approach

3.1 The Model

Compared with generative models, discriminative models can extends conveniently, they can integrate various features into the model. This paper gives a discriminative framework for bilingual news alignment based on the linear modeling approach. Within this framework, we can design various feature functions according to the bilingual news knowledge. Each feature function is associated with a feature weight. Given a Chinese news article, for every English news article candidate, we can calculate the linear combination of features as an overall score. The alignment result is the one with the highest overall score. A linear model not only allows for easy integration of new features, but also admits optimizing feature weights directly with respect to evaluation metrics.

For bilingual news alignment, we propose a linear model:

$$score(e,c) = \sum_{m=1}^{M} \lambda_m h_m(e,c) \qquad (1)$$

e and c represent an English news article and a Chinese news article respectively. $h_m(e,c)$ is a feature function with weight λ_m. For a news pair $<e,c>$, the linear combination of all features gives its overall score $score(e,c)$.

3.2 Training

Minimum error rate training (MERT) (Och, 2003) is an algorithm for optimizing parameters (i.e., feature weights) in statistic machine translation. MERT doesn't optimize parameters through maximum likelihood estimation, it tries to find parameters that result in the best F-measure or best value of other metrics directly. MERT optimizes only one parameter each time and keep all other parameters fixed. This process runs iteratively over M parameters until the overall loss on the training corpus does not decrease. Similarly, in the task of bilingual news alignment, we can use MERT algorithm to optimize the weight of each feature function with respect to the final precision of the alignments directly.

Let S be the size of the Chinese news set needed to find their alignments. For each Chinese news c_s to align to, there are K English news candidates $Cands_s=\{e_{s,1}, e_{s,1}, \ldots, e_{s,K}\}$ to align from. Parameters $\lambda_1^M=\{\lambda_1,\lambda_2,\ldots,\lambda_M\}$ are the weights of features $h_1^M=\{h_1,h_2,\ldots,h_M\}$. $e_{s,right}$ represents the right alignment of c_s tagged by human in the candidates. We use precision as the evaluation metric of bilingual news alignment, so our goal is to find a set of feature weights that maximize the precision on the training corpus:

$$\hat{\lambda}_1^M = \arg\max_{\lambda_1^M} \left\{ \sum_{s=1}^{S} \delta(\hat{e}(c_s;\lambda_1^M), e_{s,right}) \right\} \tag{2}$$

$$\hat{e}(c_s;\lambda_1^M) = \arg\max_{e\in Cands_s} \left\{ \sum_{m=1}^{M} \lambda_m h_m(e,c_s) \right\} \tag{3}$$

After given initial values and ranges of the parameters, in each iteration, MERT optimizes parameters from the first dimension to the last. For instance, when adjusting the *i-th* dimension parameter λ_i, MERT keeps the other parameters fixed. So the overall score of the alignment between $e_{s,k}$ and c_s is:

$$\begin{aligned} score(e_{s,k},c_s) &= \sum_{m=1,m\neq i}^{M} \lambda_m h_m(e_{s,k},c_s) + \lambda_i h_i(e_{s,k},c_s) \\ &= a_{s,k} + b_{s,k}\lambda_i \end{aligned} \tag{4}$$

where $a_{s,k} = \sum_{m=1,m\neq i}^{M} \lambda_m h_m(e_{s,k},c_s)$, $b_{s,k}=h_i(e_{s,k},c_s)$.

It's a linear function with parameter λ_i corresponding a line. So the set of candidates in $Cands_s$ defines a set of lines. The decision rule in Equation (3) states that $\hat{e}(c_s;\lambda_1^M)$ is the line with the highest model score for a given λ_i. The selection of λ_i for each news pair ultimately determines the precision at λ_i.

As the precision can only change if we move to a λ_i where the highest line is different than before, Och (2003) suggests only evaluating the precision at values in between the intersections that line the top surface of the cluster of lines. Figure 1 shows four candidate alignments in dimension λ_i. The upper envelope is highlighted in bold, it's constituted by the topmost line segments. The upper envelope indicates the best candidate alignments the model predicts with various values of λ_i. We just need to find the critical intersections where the topmost line changes one by one in the direction of λ_i growing, rather than all possible K^2 intersections between the K lines. We can find the closest intersection on current topmost line to current critical intersection as a new critical intersection and then update the new current topmost line and new current critical intersection to find the next critical intersection. In the interval (*leftbound,* λ_{ia}], $e_{s,2}$ has the highest score. Similarly, the best candidate are $e_{s,3}$ for (λ_{ia}, λ_{ib}], $e_{s,4}$ for (λ_{ib}, λ_{ic}], $e_{s,1}$ for (λ_{ic}, *rightbound*]. The optimal $\hat{\lambda}_i$ can be found by collecting all topmost lines related intersections on the training corpus and choosing one λ_i that results in the maximal precision value.

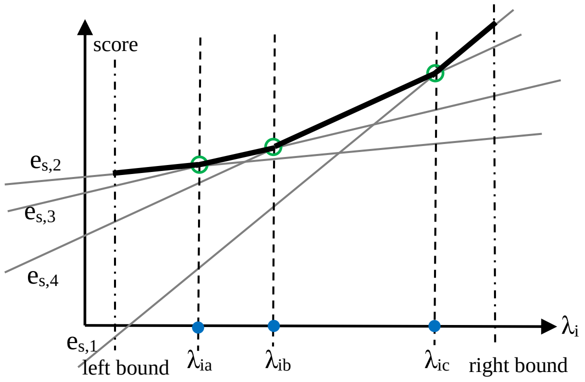

Figure 1. An example of MERT algorithm.

4 Bilingual News Alignment Features

The most significant advantage of discriminative model is to define useful features that capture various characteristics of bilingual news alignment. We can include various metrics such as bilingual vector space model as features directly. In our model, the feature set is composed of linguistic-dependent features and linguistic-independent features. We extract linguistic-dependent features such as bilingual vector space model (BVSM) cosine similarity, bilingual text graph similarity, bilingual named entity similarity, and linguistic-independent features such as news publishing date interval, content length difference. Title of news is an important characteristic too. We also use length difference, bilingual cosine similarity, and bilingual named entity similarity between news titles as features in our model. Because titles are always short, building text graphs of titles is insignificant, we don't use bilingual text graph similarity between titles as a feature in our model.

When calculating the feature values, a bilingual dictionary is used. It can be obtained after word alignment process in machine translation by a statistical machine translation toolkit GIZA++, which implements IBM models and HMM model for word alignment. This dictionary can give the translation probability of any two words in different languages.

4.1 Linguistic-Dependent Features

Bilingual Vector Space Model Cosine Similarity

VSM is a popular model for measuring similarity between monolingual documents with *tfidf* weights. Let $D=\{d_1, d_1, \ldots, d_{|D|}\}$ be a set of documents and $T=\{t_1, t_2, \ldots, t_u\}$ the set of distinct terms occurring in *D*. For term *t* in document *d*, its *tfidf* weight is defined as:

$$tfidf\,(d,t) = tf\,(d,t) \times \log\left(\frac{|D|}{df\,(t)}\right) \qquad (5)$$

Here *tf(d,t)* denotes the frequency of term *t* in *d*, *df(t)* is the number of documents in which term *t* appears, and |*D*| is the size of the document set.

Then document *d* can be represented by a vector $\overrightarrow{t_d} = (tfidf\,(t_1), \ldots, tfidf\,(t_u))^T$. For bilingual documents, let D_C and D_E denote the Chinese document set and English document set respectively. Their distinct term sets are $T_C=\{t_{c1}, t_{c2}, \ldots, t_{cm}\}$ and $T_E=\{t_{e1}, t_{e2}, \ldots, t_{en}\}$, *m* and *n* are the term numbers in T_C and T_E. The bilingual dictionary can be represented by a matrix $P_{m\times n}$, element P_{ij} in $P_{m\times n}$ is the probability of Chinese term t_{ci} being translated to English term t_{ej}.

For Chinese news d_c and English news d_e, they can be represented by vectors $\overrightarrow{t_{d_c}}$ and $\overrightarrow{t_{d_e}}$, then we can calculate the similarity between them as:

$$SIM\,(d_c, d_e) = \frac{(P^T \overrightarrow{t_{d_c}}) \cdot \overrightarrow{t_{d_e}}}{|P^T \overrightarrow{t_{d_c}}| \times |\overrightarrow{t_{d_e}}|} \qquad (6)$$

Bilingual Text Graph Similarity

Keywords can imply the topic of a text, so we assume that the similarity between bilingual news can be measured through calculating the similarity of their keywords. Rada et al. (2004) propose an innovative unsupervised graph-based model for keyword extraction. In bilingual news alignment, we build a text graph for each news article at first. Before the text graph is created, preprocessing like word segmentation on Chinese texts, stemming on English texts and Part-of-speech tagging on both sides. After moving stopwords, we choose words tagged with noun, verb and adjective as terms to build text graphs, each distinct term generates a vertex. If two terms co-occur in a fixed-size window (i.e., 3) in the origin news article text, an undirected line will be generated between the two vertexes that represent the two terms. After we run Google's PageRank (Brin and Page, 1998) on the graph, we get the weight of each vertex. The vertexes are ranked by their weights, and we extract terms whose corresponding vertexes rank in top 25% as keywords. Finally, we calculate the translation probability of the two keyword sets to evaluate the similarity between Chinese news d_c and English news d_e as:

$$p(d_c, d_e) = \frac{\sum_{i=1}^{I} w_{c_i} \sum_{j=1}^{J} p(e_j \mid c_i)}{\sum_{i=1}^{I} w_{c_i}} + \frac{\sum_{j=1}^{J} w_{e_j} \sum_{i=1}^{I} p(c_i \mid e_j)}{\sum_{j=1}^{J} w_{e_j}} \qquad (7)$$

where $\{c_1, c_2, \ldots, c_I\}$ and $\{e_1, e_2, \ldots, e_J\}$ are the sets of keywords in Chinese news d_c and English news d_e respectively. $p(e_j|c_i)$ and $p(c_i|e_j)$ are the probabilities of translating c_i to e_j and e_j to c_i, which given by the bilingual dictionary. w_{c_i} and w_{e_j} are weights of vertexes representing c_i and e_j in the two text graphs respectively.

Bilingual Named Entities Similarity

Named entity mainly refers to the terms of person names, place names or organization names. Friburger et al. (2002) and Montalvo et al. (2007) point out the effect of named entity recognition in improving monolingual and bilingual document clustering respectively. So we extract named entities through named entities recognition tools released by Stanford University, and then calculate the similarity between the two named entity sets as a feature according to Equation (7).

4.2 Linguistic-Independent Features

In bilingual news alignment, some features are linguistic-independent, but they often imply the alignment of two bilingual news. We integrate two linguistic-independent features in our model. One is news publishing date interval, the other is the ratio of the difference to the sum of the two news texts' lengths. A short interval and a small length difference ratio may imply the two news are possibly aligned to each other.

5 Experiments

5.1 Experimental Setup

The experiments were conducted on a Chinese-English corpora. The data are from two websites of two news publications: Xinhua News (Chinese, http://www.xinhuanet.com) and The New York Times (English, http://www.nytimes.com). The news are published in October 2011, and the topics contain Japan nuclear leak, Thailand floods, Mexican drug, Yemen unrest and so on. We choose 203 Chinese news and 215 English news which can represent the hot topics in the month. Table 1 shows the statistics of our reference data for bilingual news alignment. For each Chinese news, we tagged one or *n* most similar English news from the 215 English news candidates. Table 2 shows the alignments information between the bilingual news.

	news number	words	terms	size	average length
Chinese news	203	122249	10091	761KB	602
English news	215	190550	16484	1331KB	886

Table 1. Statistics on evaluation data.

n	1	2	3	4	5	6	8	9	12	total
These Chinese news number	152	17	15	3	1	6	2	3	4	203

Table 2. Alignments information of the data, n denotes the number of English news tagged aligned with one given Chinese news.

We use precision of the right alignments to evaluate our model:

$$precision=\frac{the\ number\ of\ right\ alignments}{the\ number\ of\ Chinese\ news\ in\ test\ data} \quad (8)$$

We use 4-fold cross-validation to train and test our model, and calculate the average precision as final precision of our model.

5.2 Baseline

Baseline 1: Bilingual Vector Space Model (BVSM)
We calculate the BVSM cosine similarity score between the bilingual news according to Equation (6) in section 4 and choose the candidate with the highest score as the aligning result.

Baseline 2: Machine Translation Based Method
We translate Chinese news into English by Bing's translation API (http://www.microsoft.com/en-us/translator/developers.aspx). Then we use vector space model and calculate the monolingual cosine similarity score. We choose the candidate with the highest score as the aligning result.

5.3 Results

The experiments use 4-fold cross-validation to get the average precision of our model. We also recorded the time each method cost. Table 3 shows the results.

It is worth noting that our approach results in a better precision than the two baselines and is much more efficient than the machine translation based method. The time cost in our approach is mainly spent on processes of part-of-speech tagging and named entity recognition. While machine translation based method spent more than 95% time on translation. Machine translation consumes a lot of resources and time, so MT-based method cannot meet the needs of bilingual news alignment.

Method	Precision	Cost Time (s)
Baseline1	45.84%	4
Baseline2	56.18%	168
Our Method	58.63%	13

Table 3. Performance of the methods

We also evaluated the contributions of each feature. We excluded one feature and trained the model again. The more new model declined, the more important this feature is. Table 4 shows the contribution of each feature.

Features the Model Use	Precision	Decline
ALL	58.63%	0.00%
ALL-LEN	57.62%	-1.01%
ALL-NE	55.17%	-3.46%
ALL-GRAPH	52.25%	-6.38%
ALL-BVSM	51.24%	-7.39%
ALL-DATE	51.73%	-6.90%
ALL-TitleFeatures	54.19%	-4.44%

Table 4. Contribution of each feature. ALL denotes all of the features, and LEN for length difference feature, NE for named entity feature, GRAPTH for text graph feature, BVSM for bilingual vector space model feature, DATE for news publishing date interval feature, TitltFeatures for the features related to news titles.

From Table 4 we can find that the most important features in our model are bilingual vector space model cosine similarity, news publishing date interval and text graph similarity. The length difference feature is not obvious. The title features are also important in the bilingual news alignment model. Named entity recognition can also help the task of bilingual news alignment. An example of the weights of parameters optimized by MERT algorithm is shown in Table 5.

Feature	Publish Date Interval	Content Length Difference	Content BVSM Cosine Similarity	Text Graph Similarity
Weight	-0.060	-0.067	3.504	2.737
Feature	Content Named Entities Similarity	Title Length Difference	Title BVSM cosine similarity	Title Named Entities Similarity
Weight	0.049	-0.072	2.208	0.772

Table 5. An example of the weights of parameters optimized by MERT algorithm

6 Conclusion

In this paper, we proposed a bilingual news alignment model based discriminative framework. We designed various features according to the characteristics of the news such as bilingual vector space model cosine similarity, bilingual text graph similarity, bilingual named entity similarity, news publishing date interval and the difference between the news lengths. Technologies such as Part-of-speech tagging and Named Entity Recognition are used to calculate the semantic similarities between words or text as feature values. We used minimum error rate training algorithm to optimize the feature weights. Experiments on Chinese-English news data show that our model outperforms bilingual vector space model and machine translation based method, especially more efficient than the latter.

In bilingual news alignment, our model relies on the bilingual dictionary. While many Out-of-Vocabulary (OOV) words often appear in news, which usually result in a bad precision. In future work, we will try to improve our model by using knowledge of Wikipedia and WordNet to solve OOV problem and evaluate semantic similarity between words more precisely.

Acknowledgements

This research is supported by the 863 Program under the grant No.2012AA011102 and No. 2011AA 01A207 and by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office.

References

N. Friburger and D. Maurel. 2002. Textual similarity based on proper names. In *Proceedings of Workshop on Mathematical Formal Methods in Information Retrieval at th 25th ACM SIGIR Conference.*

Anna Huang. 2008. Similarity Measures for Text Document Clustering. In *Proceedings of New Zealand Computer Science Research Student Conference (NZCSRSC).*

Jagadeesh Jagarlamudi, Hal Daume III, Raghavendra Udupa. 2011. From Bilingual Dictionaries to Interlingual Document Representations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics(ACL): shortpapers, pages 147–152.*

Yang Liu, Qun Liu, and Shouxun Lin. 2010. Discriminative Word Alignment by Linear Modeling. *Computational Linguistics, 36(3): 303-339.*

Rada Mihalcea, Courtney Corley, Carlo Strapparava. 2006. Corpus-based and Knowledge-based Measures of Text Semantic Similarity. In *Proceedings of American Association for Artificial Intelligence(AAAI).*

Rada Mihalcea, Paul Tarau. 2004. TextRank: Bringing Order into Texts. In *Proceedings of the Conference* *on Empirical Methods in Natural Language Processing (EMNLP).*

Soto Montalvo, Raquel Martinez, Arantza Casillas, Victor Fresno. 2007. Bilingual News Clustering Using Named Entities and Fuzzy Similarity. In *Matosek, V. Mautner, P. (eds.) TSD 2007. LNCS (LNAI), vol.4629, pp. 107-114. Springer, Heidelberg (2007).*

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics (ACL)*.

Bruno Pouliquen, Ralf Steinberger, Camelia Ignat, Emilia Kasper, Irina Temnikova. 2004. Multilingual and cross-lingual news topic tracking. In *COLING '04 Proceedings of the 20th international conference on Computational Linguistics Article No. 959*

Bruno Pouliquen, Ralf Steinberger, Olivier Deguernel. 2008. Story tracking: linking similar news over time and across languages. In *Proceedings of the workshop on Multi-source Multilingual Information Extraction and Summarization, pages 49-56.*

Ralf Steinberger, Bruno Pouliquen, Johan Hagman. 2002. Cross-lingual Document Similarity Calculation Using the Multilingual Thesaurus EUROVOC. In *Conference on Computational Linguistics and Intelligent Text Processing (CICLing)*.

Thuy Vu, Ai Ti Aw, Min Zhang. 2009. Feature-based Method for Document Alignment in Comparable News Corpora. In *Proceedings of the 12th Conference of the European Chapter of the ACL, pages 843-851*.

Yiming Yang, Jaime G.Carbonell, Ralf D. Brown, Robert E. Frederking. 1997. Translingual Information Retrieval:Learning from Bilingual Corpora. *In Proceedings of the 15th International Joint Conference on Artificial Intelligence (IJCAI)*.