Domain adaptation for SMT using sentence weight

Xinpeng Zhou, Hailong Cao, Tiejun Zhao

Harbin Institute of Technology, Harbin, China
{zhouxinpeng, hailong, tjzhao}@mtlab.hit.edu.cn

Abstract. We describe a sentence-level domain adaptation translation system, which trained with the sentence-weight model. Our system can take advantage of the domain information in each sentence rather than in the corpus. It is a fine-grained method for domain adaptation. By adding weights which reflect the preference of target domain to the sentences in the training set, we can improve the domain adaptation ability of a translation system. We set up the sentence-weight model depending on the similarity between sentences in the training set and the target domain text. In our method, the similarity is measured by the word frequency distribution. Our experiments on a large-scale Chinese-to-English translation task in news domain validate the effectiveness of our sentence-weight-based adaptation approach, with gains of up to 0.75 BLEU over a non-adapted baseline system.

Keywords: domain adaptation, sentence weight, Statistical Machine Translation

1 Introduction

Statistical Machine Translation (SMT) systems are trained on a large parallel corpus. In general, the sentences in a parallel training set usually come from multiple domains, such as news, laws, conference proceedings and so on. However, a translation system trained on a mix-domain corpus can't take into account the difference in translation between different domains for the same word. When fulfilling the translation tasks from all domains with such a system, the results are usually poor. Therefore, domain adaptation is crucial for SMT systems. We distinguish the domain information of the sentences in training set by adding weights, which can improve the translation quality. Recently, there are many studies about domain adaptation for SMT systems. Some researchers use a mixture model for multi-domain model adaptation (Foster and Kuhn, 2007; Koehn and Schroeder, 2007; Finch and Sumita, 2008). It is a coarse-grained approach for domain adaptation. The mixture models are composed by several language models and translate models. This method can achieve a good performance, but it can't make full use of the domain information in the training set. Some other researchers propose a new approach for domain adaptation by filtering the out-domain data (Eck et al., 2004; Foster et al., 2010; Axelrod et al., 2011). By using the in-domain data the translation quality can be improved. In this method, only

a part of the training sentences are used and they may filter out the sentences which belong to out-domain data but is helpful to improve the translation.

We describe a novel sentence-level weight method for domain adaptation in this paper. We assign weights which depend on the word distribution to each sentence pair in the training corpus. Then, we generate our domain adaptation translation system by using the sentence weights in the training process. Our method is a fine-grained approach, and we use all the sentences in the training set to train our translation system. Experimental results have shown that our method can significantly improve the translation quality on multiple domains translation task over a standard phrase-based SMT system (Koehn et al., 2003).

The rest of this paper is organized as follows: Related work on domain adaptation is presented in Section 2. The proposed approach is explained in Section 3. Experiment and results are presented in Section 4. Section 5 concludes the paper and suggests future research directions.

2 Related work

Domain adaptation is an active topic in statistical machine translation. Many researchers focused on this area. Foster and Kuhn (2007) investigated the mixture model which is consists of multiple language models and translation models by liner or non-liner interpolation. This approach can improve the translating quality, but it can't make full use of the domain information in corpus. L ü et al. (2007) used the weight of the training sentences to get an in-domain subset according to the similarity with the test data by using information retrieval models. Matsoukas et al. (2009) used a discriminative training method to estimate weights for sentences in corpus. A translation system trained with sentence weights can improve the performance. However, their method may potentially lead to over-fitting, as the characteristic function is complex and the number of parameter is large. Moore and Lewis (2010) tried to select the training data by cross-entropy. This method is helpful to improve the translation quality, but it may filter out the out-domain data which also can improve the performance. Sennrich et al. (2012) invested the translation model perplexity minimization to set model weights in mixture modeling. Banerjee et al. (2013) explored a quality estimation-guided data selection method using the target-domain data which is poorly translated. This method can improve the quality of training set, however it may also filter out the data which contribute to the performance.

In this paper, we describe a sentence weight-based domain adaptation method. Applying the sentence weights which depend on the word distribution to the training process, we can get our domain adaptation system. Our approach has the following advantages over previously mentioned techniques:

1. It is a fine-grained method and can utilize the information in the corpus flexibly compared to the mixture modeling approach.
2. Unlike the filtering data method, our system is trained on the entire training set in which each sentence has a weight. We can change the influence of each sentence in translation model by the weights.

3. We build sentence weights model by word distribution which make the calculation efficient and simple.

3 Sentence weight domain adaptation system

In this section we describe our domain adaptation translation system. We get our translation system by training the translation model with the sentence weights. In our method, the weights of training sentences are crucial to improve the performance. In our model, we use the maximum likelihood estimation to learn the weights. Our sentence-weight model is established on the monolingual corpus and we use the source language sentences in our model. The steps to calculate weights are as follows:

1. The establishment of sentence weights model. First, we assign a weight variable to each sentence. With these variables we calculate our words distribution. The frequency values are the non-linear function of weight variables. After that, we generate our model by calculating likelihood function of weights on target-domain text.
2. Model solution. The weights that we need in translation model are the values of variables when the likelihood function gains the maximum.

Our weights rely on target domain text, and they give preference to target domain. During rule extracting, we can improve the adaptation of rule by weighting training sentences. Training with the weights can both increase the adaptation of translation model and incorporate the domain information into the translation system. Our domain adaptation system is based on a phrase-based translation system and we improve the performance by training it with the sentence weights. The detailed description of the proposed approach is as follows:

First, we will introduce the variables and symbols used in this paper. We formalize the parallel training corpus like this:

$$C_{Train} = \{(f_1, e_1), (f_2, e_2), ..., (f_n, e_n)\}$$

where the (f_i, e_i) denotes the *ith* sentence pair in parallel training set, the f_i and the e_i represent the source language sentence and its translation, respectively. The subscript n is the number of the sentence pairs in the training corpus.

The target-domain text is formalized as follows:

$$C_{tar} = \{s_1, s_2, ..., s_m\}$$

where the s_i denotes the *ith* sentence in the target domain text. And the subscript m is the number of sentences in target domain corpus.

3.1 The Sentence-Weight Model

Our sentence-weight model is established on the training corpus and target-domain text according to domain characteristics. The goal of our sentence-weight model is to

make the training corpus similar to the target domain text. We estimate the weight by the similarity between the training sentences and the target-domain text. Many features can be found in a text for calculating the similarity, such as the word alignment, the lengths of sentences, and the words distribution. In our method, we find that the words distributions of different texts are similar when the texts belong to the same domains. Therefore, we use the words distribution to calculate the similarity between the training corpus and the target domain text.

First, we assign a variable to each sentence pair in the training corpus. We formalize the sentence weights as follows:

$$\boldsymbol{\lambda} = \{\lambda_1, \lambda_2, ...\lambda_i, ..., \lambda_n\}$$

where the λ_i references the weight assigned to *ith* sentence in corpus. The subscript n represents the dimensionality of the weight vector, which also is the number of parallel sentences in the training corpus.

The set of words in the training corpus is formalized as follows:

$$\mathbf{W} = \{w_1, w_2, ...w_i, ..., w_k\}$$

where w_i is the *ith* word in the words set, and the subscript k is the number of words in the corpus. The frequency of a word is calculated with all the weights. Formula 1 shows how we get our words frequency.

$$p(w_i) = \frac{\sum_{j=1}^{n} \sigma(w_i, f_j)\lambda_j}{\sum_w \sum_{j=1}^{n} \sigma(w_i, f_j)\lambda_j} \tag{1}$$

where the $\sigma(w_i, f_j)$ is a function to count the number of the word w_i in the sentence f_j. When the word w_i does not occur in the sentence f_j, the value of $\sigma(w_i, f_j)$ is zero. We obtain the word distribution by formula 1. For every word in training corpus, the frequency is a non-line function of the sentence weights.

As the weights we need are related to the target domain, we calculate the likelihood estimation of weights on the target domain text. Our likelihood estimation of weights shows as formula 2.

$$\begin{aligned} L(\boldsymbol{\lambda}) &= \prod_{j=1}^{m} \mathrm{P}(s_j) \\ &= \prod_{j=1}^{m} \prod_{i=1}^{l_i} p(w_{ji}^{'}) \end{aligned} \tag{2}$$

where the s_j represents the *jth* sentences in target domain text, and the $\mathrm{P}(s_j)$ is its probability. In our model, the sentence probability is the multiplication of words' frequencies in the sentence. The symbol $w_{ji}^{'}$ represents the *ith* word in the sentence s_j, and the $p(w_{ji}^{'})$ is the frequency of it.

In order to use the domain information, we set up it between the training corpus and target domain text. We can get weights for every domain by using the different target domain text.

3.2 The Optimal Weights

In this section we will introduce the method to get the best weights depending on the model generated in section 3.1. The weights we need in our domain-adaptation system are the values on which the sentence-weight model takes the maximum. As the number of variable is too large to calculate, we use the log-likelihood function of weights to optimize the calculation. The optimized model shows as the formula 3.

$$\log L(\boldsymbol{\lambda}) = \sum_{j=1}^{m}\sum_{i=1}^{l_j}\log p(w_{ji}^{'}) \tag{3}$$

So, the weight values we need are like this:

$$\begin{aligned}\tilde{\boldsymbol{\lambda}} &= \arg\max_{\boldsymbol{\lambda}}\sum_{j=1}^{m}\sum_{i=1}^{l_j}\log p(w_{ji}^{'}) \\ &= \arg\max_{\boldsymbol{\lambda}}\sum_{j=1}^{m}\sum_{i=1}^{l_j}\log\frac{\sum_{l=1}^{n}\sigma(w_{ji}^{'}, f_l)\lambda_l}{\sum_{w}\sum_{l=1}^{n}\sigma(w_{ji}^{'}, f_l)\lambda_l}\end{aligned} \tag{4}$$

In a SMT system, the number of sentences in the training corpus usually is huge. So we should deal with a large number of variables in our model. It is a hard task to get the best values when the problem has a lot of variables. In our method, we use the L-BFGS algorithm (Sennrich, 2012) to deal with the sentence-weight model. We can get an approximate optimal solution by the L-BFGS algorithm.

3.3 Domain Adaptation Translation System

The domain adaptation translation system is based on a phrase-based machine translation system. By altering the phrase translation table, the performance of a translation system can be improved. We apply the weights in the training process to calculate the translation probability of a phrase pair. In a phrase-based translation model, the translation probability of the phrase pair $(f^{'}, e^{'})$ is defined as follows:

$$\phi(f^{'} \mid e^{'}) = \frac{count(f^{'}, e^{'})}{\sum_{f^{'}} count(f^{'}, e^{'})} \tag{5}$$

where the $count(f^{'}, e^{'})$ is the occurrence number of phrase pair $(f^{'}, e^{'})$. The phrase pairs are extracted form a sentence pair with the word alignment. In our domain adaptation system, we use the weights to calculate the translation probability of a phrase pair. Our method is like this:

$$\phi(f^{'} \mid e^{'}) = \frac{\sum_{i=1}^{n} \Phi(f^{'}, e^{'} \mid f_i, e_i)\lambda_i}{\sum_{f^{'}} \sum_{i=1}^{n} \Phi(f^{'}, e^{'} \mid f_i, e_i)\lambda_i} \tag{6}$$

where $\Phi(f^{'}, e^{'} \mid f_i, e_i)$ is the occurrence number of phrase pair $(f^{'}, e^{'})$ in sentence pair (f_i, e_i), and λ_i is the weight of this sentence pair.

4 Experiments and Result Analysis

4.1 Data

We evaluate our domain adaptation approach on Chinese-to-English machine translation task. The training corpus in our experiment is a mixture corpus consisting of sentences from several domains such as news, laws, conference proceedings and so on. The target domain is the news in our experiment. The language model is training on the target side of FBIS and Gigaword. The Chinese-English parallel corpus that we use in our experiment is released by LDC[1]. The detail of corpus we used is shown in Table 1.

Table 1. The corpus. In the *role* column: train=train set, dev = development set, test=test set, tar=target-domain set; In the *genres* column: cp= conference proceedings, nw= newswire.

role	corpus	genres	sent
training	LDC-Hong Kong Hansards	cp	1,297k
	LDC-Hong Kong Laws	laws	400K
	LDC-Hong Kong News	nw	702k
	FBIS+ Gigaword	nw	12,701k
dev	NIST 2002 OpenMT Evaluation	nw	878
test	NIST 2005 OpenMT Evaluation	nw	1082
tar	LDC-Chinese news	nw	11,795k

In our experiment, the training data consist of two parts: the monolingual data (FBIS+ Gigaword) and the bilingual data (Hong Kong Hansards, Hong Kong Laws, Hong Kong News). The monolingual data is used to train the language model, and the bilingual data is used to train the translation model. The target domain text is a monolingual data set, and it is used in the sentence-weight model. The number of Chinese-English parallel sentences is about 3.37 million and the number of sentences in Chinese News is about 1.7 million. The language training data contains about 12.7 million newswire sentences.

4.2 System Description

Our baseline is a standard phrase-based SMT system. Given a source sentence, it can find the most likely translation according to the phrase translation table and the

[1] LDC2003T05, LDC2004T08, LDC2005T06, LDC2010T10, LDC2010T14, LDC2011S01

Viterbi approximation. Our domain adaptation system consists of two components, the sentence-weight model and the domain adaptation translation system.

In the sentence-weight model, we generate the likelihood estimation of weights by calculating the probabilities of the target domain sentences. During this process, we use the Laplace smoothing (Field, 1988) to deal with the unknown words, and we also remove the stop words whose frequency is in the top 100 of words distribution. We use the L-BFGS algorithm (1989) to get the optimal value. In this algorithm, the initial values of all weights are 1.0.

Our translation model are trained on the bilingual data. We use the GIZA++(Och and Ney, 2003) to align the words in the bilingual sentence pair in both directions. Our 4-gram language model are trained on a in-domain monolingual corpuswith modifiedKneser-Ney smoothing(Kneser and Ney, 1995) through the SRILM language modeling toolkit(Stolcke, 2002). The evaluation metric we used for the translation quality is the BLEU4 (Papineni et al., 2002).

4.3 Result

The performance of our domain adaptation trainalation system is shown in table 2. In the experiment, the target domian is news, we test our system on the same multiple domains corpus as the baseline system. As a comparison, we also trained a phrase-based translation system only with the news domain training data.

Table 2. The weight-based translation sysytem result. In the *weight* column: PBMT = the baseline system, PBMT+W = sentence-weight-based domian adaptation translnation system, PBMT+D= the phrase-based translation system trained with in-domain training data.

system	BLEU
PBMT	26.90
PBMT+W	27.65
PBMT+D	27.73

Table 2 shows that for the same translation task from the news test set, the baseline system get a score of 26.90, and the score of the in-domain translation systemand is 27.73. The score of our domian adaptation translation system is 27.65, the gain is 0.75 according to the baseline. Through the table 2 we can see that our system trained with the sentence weights can get a better performance over the baseline.

4.4 Analysis

In the experiment we random sample five values from the weights vector to check the performance of our sentence-weight model. Table 3 list several target domain sentences and five training sentences with their corresponding weights. In the table 3, according to the content of the sentences we can know that the fifth sentence in training sample set has the most number of the same words with the target sentences. Therefore, in our method the fifth sentence has the greatest similarity among the five training sentences. And the next two are the first sentence and the fourth sentence in

order of the similarity with the target sentences. In the weight column, the values of weights are, in order, the fifth, the first, the fourth, the second and the third.

Table 3. The sentence weights sample result.

role	weight	sentence
Target	--	国际足联规定 , 从今年 7 月 1 日 开始 , 球员的背心上不 得书写任何文字 。
	--	香港居民在沪申请设立个体工商户将享受绿色通道.
	--	民政部处罚中国地区开发促进会：停止活动 3 个月。
	--	按国家艾滋病统计数据，北京市艾滋病病毒感染者的报告数，在全国排名第 8 位。
Training	0.9864	当局亦已答应优先检讨亲父鉴定诉讼条例, 以减少根据该条例规定提出申请的母亲的不便。
	0.9684	现在付诸表决, 赞成的请举手。
	0.8658	(1)除第(2)款另有规定外, 任何人的申索如已被根据第 113 条不准予或只局部准予。
	0.9765	民政事务总署现已在各区开放共十四个临时避寒中心, 供有需要的市民避寒。
	1.074	他说, 在香港有三人获证实死于该病毒, 另有两名患有重病的病人, 其死因可能亦与该病毒有关。

The data in the table 3 shows that the weight of sentence with greater similarity has a larger value. From table 3 we can know that our sentence-weight model can increase the weight value of sentence with high similarity to the target domain text and reduce the weight value of sentence with low similarity.

In our domain adaptation translation system, we improve the performance by increasing the adaptation of translation model through the sentence weights. During the experiment we sample the phrase translation rules in our translation system, the baseline system and the news domain translation system. Table 4 lists some difference between our system and other translation systems in the phrase translation rules.

Table 4 shows the difference in the translation rules between our system and compared system. The phrases in the *rules* column in table 4 are the translations for the phrases in the *source* column. For the phrase ”成为重要”, the probability of translation rules which translate the source phrase to ”become an important” and ”a leading” are all 0.6 in baseline, while we change the probabilities of rules by the sentence weights. Our system improves the probability of the translation rule which translate to “become an important” to 0.633204 and reduce the probability of the other one to 0.598505. In decoding step, the best translation to the phrase ”成为重要” is ”become an important” in our system, while the baseline choose the phrase ”a leading”. For the phrase ”促进社会稳定” , our system also change the probability of translation rules by weights, and when decode for this phrase, the result ”the promoting social stability” is the best translation, while the baseline translate it to ”foster social stability”. From he table 4 we can know that the translation rules of our domain adatptation system are more similar to that of the news domain translation system. So, by using our sentence weights we can change the probability of the translation rules

and decoding result, which will lead to the improvement of the ability of domain adaptation.

Table 4. The difference in phrase translation rules between our system and other systems. In the *PBMT*, the *PBMT+W* and the *PBMT+D* column, the values are the probabilities of translation rules in the baseline system, our translation system, and the news domain system, respectively. The superscript "*" are the marks which means the translation rule is the best rule in decoding for the source phrase in the system.

source	rule	PBMT	PBMT+W	PBMT+D
成为重要	become an important	0.600000	0.633204*	0.628895*
	a leading	0.600000*	0.598505	0.600000
	an important	0.600000	0.600064	0.600000
	as an important commercial	0.200000	0.191113	0.200000
促进社会稳定	promoting social stability	0.600000	0.616564*	0.633168*
	foster social stability	0.428571*	0.395946	0.389746
	our society can be made stable	0.600000	0.608138	0.630000
	as to foster social stability	0.200000	0.197310	0.621658

The experiment result shows that our domain adaptation can improve the system performance. We sample several sentences whose translate results are different between our system and the baseline to check the translation quality. Table 5 show the translate results for the same sentences translated by our system, the baseline system and the domain adaptation system.

Table 5. The difference between the translation result. In system column: task = translation task, PBMT =baseline, PBMT+W=our domian adaptation system, PBMT+D=news domain adaptation system, ref = reference translation. In the result, we use shift characters to represent the spical character, " ' " = " ' ", " "" =" " ".

role	sentences
task	1. 明年日本经济成长速度可能放缓.
	2. 世界各国领袖陆续走访斯里兰卡视察灾情.
PBMT	1. next year, the economic growth is likely to slow the pace of japan.
	2. the world leaders continue to inspect the disaster to visit sri lanka.
PBMT+W	1. the speed of economic growth next year might be slow-down in japan.
	2. world leaders have visited sri lanka to inspect the disaster.
PBMT+D	1. japan 's economic growth next year is likely to slow speed.
	2. the world leaders to inspect the disaster will visit sri lanka.
ref	1. japan 's economy may slow down next year.
	2. world leaders tour sri lanka to assess tsunami damage.

There are two translation instances in the table 5, we list out the source sentences, the results translated by baseline system, our domain adaptation system and the news domain translation system, and the reference translation. For the instance 1, the difference between our system and the baseline is the target to slow down, in our system the target is the economic growth, while in the baseline system the target is the japan. And the meaning of the translation of our system is the same to that of the news do-

main translation system. In the sentence 1, the meaning of source sentence is that the economic growth may slow down, so, the translation of baseline is wrong and ours is accurately. When translate the instance 2, the difference in results is the translation of phrase ”走访斯里兰卡视察灾情”, our system translate it exact, while the translation of the baseline is wrong, as it turn over the order of ”inspect the disaster” and ”visited sri lanka”. In the two instances, the meaning of our system translation result is the same as that of the news domain translation system. Through the two instances we can know that by applying the sentence weights, we can improve the performance of a translation system in the target domain.

5 Conclusion

In this paper, we describe a method to estimate the sentence weights to enhance the ability of domain adaptation and improve the performance of the translation system. Firstly, we build our sentence-weight model by using the word frequency distribution. And we use the L-BFGS algorithm to get the sentence weights according to target domain text in the model. Then, we train our translation model with sentence weights and get the domain adaptation translation system. Experiment results show that our approach brings a better performance in target domain over the phrase-based translation system (Koehn et al., 2003).

Our method is a fine-grained and sentence-level domain adaptation method in machine translation. And it is also a general domain adaptation approach. Our sentence-weight model depends on the word frequency distribution, we may also generate it by other features. In future work, we will try to use other characteristics to generate the sentence-weight model to improve the translation result.

6 References

1. George Foster, Roland Kuhn. Mixture-model adaptation for SMT[J]. 2007. Workshop on Statistical Machine Translation Association for Computational Linguistics. Workshop on Statistical Machine Translation, Association for Computational Linguistics.
2. Koehn P, Schroeder J. Experiments in domain adaptation for statistical machine translation [C]//Proceedings of the Second Workshop on Statistical Machine Translation. Association for Computational Linguistics, 2007: 224-227.
3. Finch A, Sumita E. Dynamic model interpolation for statistical machine translation[C]// Proceedings of the Third Workshop on Statistical Machine Transla- tion. Association for Computational Linguistics, 2008: 208-215.
4. Zhao B, Eck M, Vogel S. Language model adaptation for statistical machine translation with structured query models[C]//Proceedings of the 20th international conference on Computational Linguistics. Association for Computational Linguistics, 2004: 411.
5. Foster G, Goutte C, Kuhn R. Discriminative instance weighting for domain adaptation in statistical machine translation[C]//Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2010: 451-459.

6. Axelrod A, He X, Gao J. Domain adaptation via pseudo in-domain data selection[C]// Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2011: 355-362.
7. Koehn, Philipp, Franz Josef Och, and Daniel Marcu. Statistical phrase-based translation[C] //Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1. Association for Computational Linguistics, 2003: 48-54.
8. L ü Y, Huang J, Liu Q. Improving Statistical Machine Translation Performance by Training Data Selection and Optimization[C]//EMNLP-CoNLL. 2007: 343-350.
9. Matsoukas S, Rosti A V I, Zhang B. Discriminative corpus weight estimation for machine translation[C]//Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2. Association for Computational Linguistics, 2009: 708-717.
10. Moore R C, Lewis W. Intelligent selection of language model training data[C]//Proceedings of the ACL 2010 Conference Short Papers. Association for Computational Linguistics, 2010: 220-224.
11. Sennrich R. Perplexity minimization for translation model domain adaptation in statistical machine tra- nslation[C]//Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2012: 539-549.
12. Banerjee P, Rubino R, Roturier J, et al. Quality estimation-guided data selection for domain adaptation of smt[J]. MT Summit XIV: proceedings of the fourteenth Machine Translation Summit, 2013: 101 -108.
13. Liu D C, Nocedal J. On the limited memory BFGS method for large scale optimization[J]. Mathematical programming, 1989, 45(1-3): 503-528.
14. Field D A. Laplacian smoothing and Delaunay triangulations[J]. Communications in applied numerical methods, 1988, 4(6): 709-712.
15. Och F J, Ney H. A systematic comparison of var- ious statistical alignment models[J]. Computational li- nguistics, 2003, 29(1): 19-51.
16. Kneser R, Ney H. Improved backing-off for m-gram language modeling[C]//Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 Internatio- nal Conference on. IEEE, 1995, 1: 181-184.
17. Stolcke A. SRILM-an extensible language modeling toolkit[C]//INTERSPEECH. 2002.
18. Papineni K, Roukos S, Ward T, et al. BLEU: a method for automatic evaluation of machine translation [C]//Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics, 2002: 311-318.