Domain Adaptation for Paraphrasing

Abstract

Paraphrases are typically tailored towards general domain of texts. Paraphrase extraction for a target domain can lead to poor phrase coverage and low quality paraphrases due to the lack of in-domain data. In this paper, we developed a framework for domain-specific paraphrasing, which applies cross-entropy difference subsampling introduced by Moore and Lewis (2010) to a multilingual paraphrase acquisition scheme. We compare the paraphrase quality for multiple domains and empirically demonstrate that domain-biased paraphrases resulting from our method is substantially more relevant to and contains more coverage of its source domain. Experimental results also indicates an improvement in the substitution task with domain-adapted paraphrases.

1 Introduction

Paraphrasing is a method of re-expressing a phrase that preserves the original meaning, usually within a certain context. Data-driven paraphrase acquisition approaches can be grouped into several categories depending on the data used for training. Monolingual paraphrasing methods uses statistical characteristics extracted from monolingual resources, such as dependency path similarities or distributional co-occurrence information (Lin and Pantel, 2001; Pasca and Dienes, 2005). Bilingual paraphrasing techniques extract paraphrase candidates by grouping English phrases that share common foreign translations in parallel corpora (Bannard and Callison-Burch, 2005). Additional paraphrase extraction methods fall between the two extremes, such as utilizing multiple English translation of the same foreign text or drawing knowledge from statistical machine translation to monolingual resources (Barzilay and McKeown, 2001; Pang et al., 2003; Quirk et al., 2004).

Existing work in paraphrase acquisition has focussed on the open-domain setting. This leads to results that are subjectively intuitive, and sometimes useful in unconstrained settings, but are potentially less useful when working in a specific domain such as chemistry, literature or art. Consider the biology domain: morphological variants of the word *divide* may often be synonymous with variants of *multiply*, such as in: *cell division*, and *cell multiplication*. This ambiguity is unlikely to hold in the mathematics domain, or even in general text. Simply increasing the training corpus size for domain-specific paraphrasing is infeasible, since data-driven paraphrase acquisition techniques require large amount of training data and prepackaged in-domain training data is hard to come by in most domains.

We construct a framework of domain adaptation for paraphrasing by first intelligently subsampling a domain-specific parallel corpus from a much larger general domain of text using a metric developed by Moore and Lewis (2010). We then apply monolingual similarity scores to the paraphrase candidates generated based on the domain-specific subsampled corpus using the pivot-based bilingual procedure. We show that the domain-biased paraphrases generated by our method are more relevant to the target domain than without the domain-adapted subsampling. In addition, experimental results illustrate that domain adaptation positively impacts the substitution quality of paraphrases.

2 Related Work

2.1 Domain Adaptation for SMT

Daumé III et al. (2010) proposed a domain adaptation approach for SMT in a fully supervised setting. For an existing learning a function that maps vectors from an input space to a decision/output space, their approach expands each dimension in the input feature space (K + 1)-fold, where K is the number of domains. The training procedure which optimizes its performance on target (in-domain) data takes into account the additional information about the domain from which the data is collected. This paradigm trades additional domain information for increased computational burden due to the increase of feature space. Moreover, this adaptation procedure is implicitly restricted to certain structures of end-to-end tasks since it is incorporated directly into the machine translation task for both training and testing. These reasons makes such adaptation inapplicable for our paraphrasing framework.

Schroeder (2007) carried out experiments to compare end-to-end machine translations performance in terms of translation scores (BLEU) for a range of domain adaptation techniques applied to training data selection. Some of the techniques include in-domain language model, interpolating language model to bias towards specific domains, and simply combining training data. They showed that best BLEU scores are achieved with by adapting the training data to the domain of interest using language models.

Yarowsky (1995) introduced the notion of *one sense per discourse* to WSD, where a word or phrase in a given document, or perhaps a collection of related documents, will tend to have just a single sense. This suggests that in order to extract paraphrases that avoid the problems of WSD, we might simply restrict ourselves to parallel corpora (when using bilingual pivot-based methods) that fall in the domain of interest. However, it is not obvious where one would naturally find large collections of parallel data, precompiled according to specific domains.

We therefore propose to automatically induce *domain-biased* parallel corpora, through extending the method proposed by Moore and Lewis (2010) for building custom training corpora for MT. In the context of language model training, Moore and Lewis (2010) compared data selection methods such as in-domain cross-entropy scoring, similar to those used in Schroeder (2007), Klakow's method (Rajasekaran et al., 2005) and cross-entropy difference scoring. They showed that difference in cross entropy results in the lowest perplexity in the test set of a specific target domain. The details of their model elaborated in Section 3.1.

In the first stage of our domain-bias paraphrasing framework, we begin with constructing an in-domain statistical language model with a monolingual collection of domain specific content. A separate ”general-domain” language model is then constructed for a collection made up of uniformly at random sampled material from the English side of a large collection of mixed domain parallel text. Each sentence in the remaining parallel corpus is scored based on the difference in cross entropy (diffCE) from the language models. Domain-biased subsampling retains sentences that satisfy certain threshold on the diffCE. The subsampled parallel corpus is subsequently fed into a paraphrase extraction pipeline.

2.2 Paraphrase Acquisition from Bitexts

Bannard and Callison-Burch (2005) proposed identifying paraphrases by pivoting through phrases in a bilingual parallel corpora. Figure 1 illustrates their paraphrase extraction process. The *target* phrase, e.g. *thrown into jail*, is found in a German-English parallel corpus. The corresponding foreign phrase (*festgenommen*) is identified using word alignment and phrase extraction techniques from phrase-based statistical machine translation (Koehn et al., 2003). Other occurrences of the foreign phrase in the parallel corpus may align to a distinct English phrase, such as *jailed*. As the original phrase occurs several times and aligns with many different foreign phrases, each of these may align to a variety of other English paraphrases. Thus, *thrown into jail* not only paraphrases as *jailed*, but also as *arrested*, *detained*, *imprisoned*, *incarcerated*, *locked up*, and so on. Bad paraphrases, such as *maltreated*, *thrown*, *cases*, *custody*, *arrest*, and *protection*, may also arise due to poor word alignment quality and other factors.

Bannard and Callison-Burch (2005) defined a paraphrase probability to rank these paraphrase candidates, as follows:

$$\begin{aligned} \hat{e_2} &= \arg\max_{e_2 \neq e_1} p(e_2|e_1) && (1) \\ p(e_2|e_1) &= \sum_f p(e_2, f|e_1) && (2) \\ &= \sum_f p(e_2|f, e_1) p(f|e_1) && (3) \\ &\approx \sum_f p(e_2|f) p(f|e_1) && (4) \end{aligned}$$

where $p(e_2|e_1)$ is the paraphrase probability, and $p(e|f)$ and $p(f|e)$ are translation probabilities from a statistical translation model.

Anecdotally, this paraphrase probability sometimes seems unable to discriminate between good

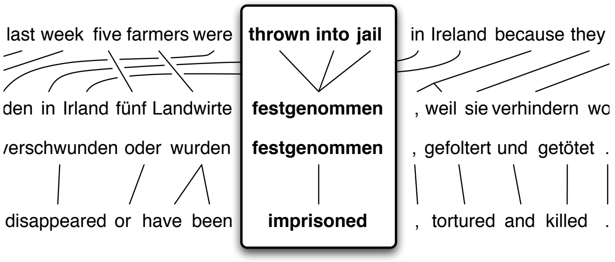

Figure 1: Using a bilingual parallel corpus to extract paraphrases.

and bad paraphrases, so some researchers disregard it and treat the extracted paraphrases as an unsorted set (Snover et al., 2010). Callison-Burch (2008) attempts to improve the ranking by limiting paraphrases to be the same syntactic type.

We attempt to rerank the paraphrases using other information. This is similar to the efforts of Zhao et al. (2008), who made use of multiple resources to derive feature functions and extract paraphrase tables. The paraphrase that maximizes a log-linear combination of various feature functions is then selected as the optimal paraphrase. Feature weights in the model are optimized by minimizing a *phrase substitution error rate*, a measure proposed by the authors, on a development set.

2.3 Monolingual Distributional Similarity

Prior work has explored the acquisition of paraphrases using distributional similarity computed from monolingual resources, such as in the DIRT results of Lin and Pantel (2001). In these models, phrases are judged to be similar based on the cosine distance of their associated context vectors. In some cases, such as by Lin and Pantel, or the seminal work of Church and Hanks (1991), distributional context is defined using frequencies of words appearing in various syntactic relations with other lexical items. For example, the nouns *apple* and *orange* are contextually similar partly because they both often appear as the object of the verb *eat*. While syntactic contexts provide strong evidence of distributional preferences, it is computationally expensive to parse very large corpora, so it is also common to represent context vectors with simpler representations like adjacent words and n-grams (Lapata and Keller, 2005; Bhagat and Ravichandran, 2008; Lin et al., 2010; Van Durme and Lall, 2010). In these models, *apple* and *orange* might be judged similar because both tend to be one word to the right of *some*, and one to the left of *juice*.

Here we calculate distributional similarity using a web-scale n-gram corpus (Brants and Franz, 2006; Lin et al., 2010). Given both the size of the collection, and that the n-grams are sub-sentential (the n-grams are no longer than 5 tokens by design), it was not feasible to parse, which led to the use of n-gram contexts. Here we use adjacent unigrams. For each phrase x we wished to paraphrase, we extracted the context vector of x from the n-gram collection as such: every (n-gram, frequency) pair of the form: (ax, f), or (xb, f), gave rise to the (feature, value) pair: $(w_{i-1}{=}a, f)$, or $(w_{i+1}{=}b, f)$, respectively. In order to scale to this size of a collection, we relied on Locality Sensitive Hashing (LSH), as was done previously by Ravichandran et al. (2005) and Bhagat and Ravichandran (2008). To avoid computing feature vectors explicitly, which can be a memory intensive bottleneck, we employed the online LSH variant described by Van Durme and Lall (2010).

This variant, based on the earlier work of Indyk and Motwani (1998) and Charikar (2002), approximates the cosine similarity between two feature vectors based on the Hamming distance in a reduced bitwise representation. In brief, for the feature vectors $\vec{u}$, $\vec{v}$, each of dimension d, then the cosine similarity is defined as: $\frac{\vec{u}\cdot\vec{v}}{|\vec{u}||\vec{v}|}$. If we *project* $\vec{u}$ and $\vec{v}$ through a d by b random matrix populated with draws from $N(0, 1)$, then we convert our feature vectors to *bit signatures* of length b, by setting each bit of the signature conditioned on whether or not the respective projected value is greater than or equal to 0. Given the bit signatures $h(\vec{u})$ and $h(\vec{v})$, we approximate cosine with the formula: $\cos(\frac{D(h(\vec{u}),h(\vec{v}))}{b}\pi)$, where $D()$ is Hamming distance.

3 Domain Adaptation

3.1 Data Selection based on Difference in Cross Entropy

As briefly introduced in Section 2, Moore and Lewis (2010) proposed an approach for subsampling of a non-domain specific corpus based on difference in cross entropy with respect to in-domain and general-domain language models. The goal of their work was to exploit a large corpus efficiently in terms of computational resources to enhance for specific machine translation tasks. In their method, a substantially larger general-domain corpus is randomly subsampled to match the size of the much smaller in-domain text. Separate language models are trained with the in-domain text and smaller general-domain corpora, with the assumption that enough data was used to train an effective language model. The cross entropy of individual sentences in the remaining general-domain corpus are calculated according to the language models.

Since a smaller value in cross entropy implies the closer resemblance of a sentence to the domain of the language model, the difference in cross entropy can serve as metric for selecting a subset of the non-domain-specific corpus that matches more with the in-domain data than the out-of-domain data. Moore and Lewis showed that, across a range of subsampled data sizes, their subsampling method produced much smaller perplexity for in-domain test data as compared to methods such as Klakow's method, in-domain cross-entropy scoring and random selection.

In our work, we utilize this data selection method as a means to bias the relevance of our paraphrases to a domain of interest. The overall domain adaptation procedure is illustrated in Figure 1. After the in-domain and out-of-domain language models are trained with the respective training data, they are used to calculate the cross entropy for the sentences in rest of the out-of-domain text. For a particular sentence s, the difference in cross entropy, defined as $\Delta\text{H}(s) = \text{H}_{IN}(s) - \text{H}_{OUT}(s)$, is used to categorize it into one of N bins, which all contain roughly the same number of sentences and are specified by the range of cross entropy difference values. The larger a $\Delta\text{H}(s)$ value is, the closer to the in-domain text this sentence s is.

The corresponding translation of the sentences in the top bins are retrieved from the out-of-domain parallel corpus and result in a domain-adapted parallel corpus. This subsampled text then undergoes the standard bilingual pivoting framework for extracting a collection of phrases and the associated domain-adapted paraphrases.

There is a trade-off between the amount of data for paraphrase extraction and the relevance of the paraphrases to the in-domain text. A larger number of top bins used for constructing a paraphrase table would lead to more coverage and confidence of the paraphrases, but the paraphrases would also be less applicable to the specific domain of interest.

3.2 Domain Adaptation for Paraphrasing

Each stage in our paraphrase framework depends on different data statistics from its the training data, we therefore investigated the effectiveness of paraphrase biasing by applying domain adaptation independently to pivot-based paraphrase acquisition and monolingual reranking.

For bilingual paraphrase acquisition, by biasing the domain of the parallel corpus from which grammar is extracted, the phrase table would potentially be dominated by content with more frequent uses in that specific domain. Hence, the resulting coverage and probability of paraphrases after bilingual pivoting are both expected to be closer to the domain of interest.

The similarity score based on monolingual distributional similarity (MonoDS) was originally proposed to be computed using the Google ngram corpus (Lin et al., 2010), which comprises of statistics from general domain of resources on the internet. While large amount of data might lead to a better vector space model for the MonoDS score, adapting a training corpus to the target domain can potentially enhance the relevance of the paraphrases reranking.

4 Experimental Setup

In this section, the details of the experimental setup are presented with respect to the diagram in Figure 2.

4.1 Data and Language Models

@@ data: in-domain, out-domain @@ srilm model params @@ tokenization, normalization, cleaning, all preprocessing (see fileInfo.txt notes and refer to first paragraphs of Moore's section 3) @@ mention only comparison between diffCE and random subsampling are done for size Europarl because of engineering/computational limitation after many attempts. But this is also precisely what we emphasize in the paper: under finite resources and time, @@ thresholding in BiP and MonoDS scores @@ averaging in analysis: mean sent, median turker @@ emphasize what's used as ”in-domain test set” and

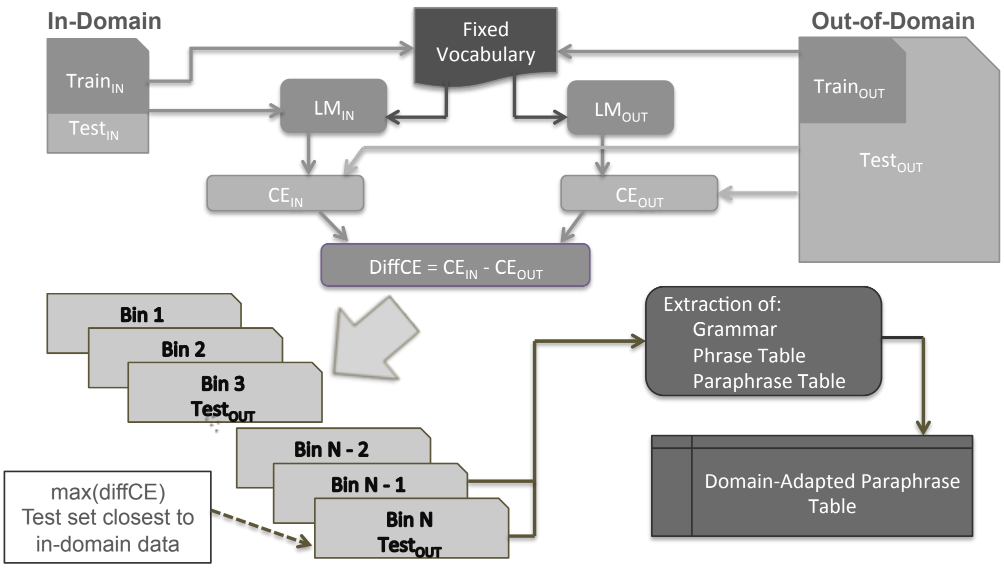
In-Domain
Fixed Vocabulary
Out-of-Domain
Train_{IN}
Test_{IN}
LM_{IN}
LM_{OUT}
$\text{Train}_{\text{OUT}}$
Test_{OUT}
CE_{IN}
CE_{OUT}
$\text{DiffCE} = \text{CE}_{\text{IN}} - \text{CE}_{\text{OUT}}$
Bin 1
Bin 2
Bin 3 Test_{OUT}
Extraction of: Grammar Phrase Table Paraphrase Table
Bin N - 2
Bin N - 1
Bin N Test_{OUT}
max(diffCE) Test set closest to in-domain data
Domain-Adapted Paraphrase Table

Figure 2: Diagram of the domain adaptation procedure based on difference in entropy. LM, CE, diffCE stand for language model, cross entropy, and difference in cross entropy, respectively. Bins of out-of-domain test data are ordered based on their diffCE values, which is proportional to the ”closeness” of each individual sentences to the targeted domain. N, denoting the number of bins, is 8 in our experiment.

will be used repeatedly for quantitative evaluation in subsequent sections

We defined our target domain through combining the biology textbook and the GENIA database() of similar sizes. The GENIA database is a collection of 1.999 biomedical publication abstracts, whereas the biology textbook is intended for generic introductory courses in biology.

A list of fixed vocabulary extracted from the unigrams in the in-domain training data and the textbook index, were used to train both of the language models. Preprocessing on the training data was performed in order to eliminate lines with non-sentence like structure or with too many special or repeated characters. Unigrams from the biology textbook was required to have repeated occurrences in order to be included in the vocabulary list.

GigaFrEn, a French-English parallel corpus verging on 1 billion words (Callison-Burch et al., 2009) was used as the general corpus from which the domain-specific corpus is built. It was constructed by conducting a large-scale web crawl targeting bilingual web sites. It captures a wide variety of topics[1], and is therefore perfectly suited for domain-adaptation studies for bilingually derived paraphrases.

The non-domain-specific data consists of the Giga-French-English and the Europarl French-English corpora, which sum up to roughly 0.6 billion words on the English side. The language models for cross entropy calculations were trained separately on data set of similar sizes to allow for comparable results.

4.2 Language Model Training for In-Domain Text

4.3 Paraphrase Extraction

@@ take content from GEMS Sect 3.3? The parallel corpus The bilingual pivoting

@@ examples

@@ 20111117b.rtf @@@

[1] These sites came from a variety of sources including the Canadian government, the European Union, the United Nations, Amnesty International, the World Health Organization, and other international organizations. The crawl yielded on the order of 40 million files, consisting of more than 1TB of data. It represents the largest and most diverse parallel corpus available for research purposes.

DiffCE		Random	
Line count	**Token count**	**Line count**	**Token count**
1,271,750	27,831,706	1,270,000	27,852,488
2,543,466	56,384,338	2,540,000	55,733,642
5,086,899	119,776,920	5,080,000	111,478,225
10,173,762	263,492,323	10,160,000	222,917,055

Table 1: Statistics of corpora used for domain adaptation

	Line/Token count	PPL$_{InDomain}$
DiffCE	573,659 / 12,779,577	319.2
Subsampling$_{Joshua}$	573,659 / 10,667,535	604.4

Table 2: Statistics of corpora used for domain adaptation method comparison and the respective in domain *test* data perplexities (PPL$_{InDomain}$)

5 Experimental Results

5.1 Subsampling Methods Comparison

Although domain-biased subsampling is expected to outperform random sampling, the common notion that ”more data beats better algorithms” implies that the effect of domain-bias can potentially be masked by the size of a randomly subsampled corpus. To investigate the gain of applying domain adaptation to paraphrasing, corpora at various sizes are subsampled with each method and used for language model training independently. The line and token counts of each pair of corpora at different sizes are listed in Table 1.

Due to computational constraints, research in MT typically employs simple subsampling methods in order to reduce the amount of data for actual processing. For example, a standard subsampling method based on n-gram overlap with a reference text for sentence selection is provided by Joshua (Li et al., 2009), an open source decoder for statistical MT. In order to compare this subsampling method with that based on difference in cross entropy, the corpus of similar size to the Joshua subsampling was extracted based on diffCE. Table 2 tabulates the corpora sizes and the perplexity for in-domain text data. The smaller perplexity produced by diffCE-based method validates cross entropy from two contrasting domains is more informative than basic n-gram overlap methods.

	N-gram				
Corpus	**1**	**2**	**3**	**4**	**Total**
Europarl	11,536	55,373	42,540	13,851	123,300
Top	16,771	91,697	67,737	18,292	194,497
Bottom	9,022	25,835	11,575	2,008	48,440
Random	14,606	66,656	45,207	11,616	138,085
Textbook	21,422	212,408	433,425	512,515	1,179,770

Table 3: Number of unique N grams in biology textbook that appear in the paraphrase tables generated with different subsampling methods, where *Top* and *Bottom* refer to the domain adapted domains closest and furthest from the domain of interest, respectively

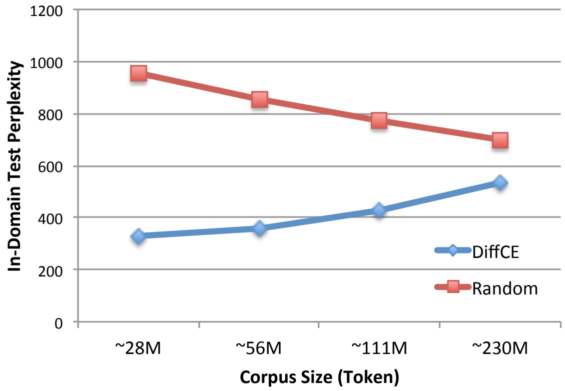

Figure 3: In-domain test data perplexity according to language models trained based on difference in entropy (DiffCE) and random sampling (random) as a function of sample size.

Table 3 reports the overlap between the unique phrases of up to 4-gram from the biology textbook and each of the paraphrase tables generated from different subsampled corpora. Although all of the paraphrase tables were constructed with corpora of very similar sizes, the domain-biased paraphrase table, labeled as *Top*, contains the most number of phrases in the biology textbook across all 4 n-gram lengths. This highlights that intelligent selection of corpus for bilingual pivoting facilitates an increase of the paraphrase coverage for a task of interest.

Domain adapting and random subsampling are compared in Figure 3 at various sizes as listed previously in Table 1. The domain specific subsampling results in less perplexity than random selection throughout the two curves, which indicates that, at each of the increasing corpus sizes, the language model trained on the data subsampled with difference in cross entropy remains better at prediction than the baseline of random sampling. This plot is comparable to the empirical result obtained in Moore and Lewis (2010), which suggested that diffCE is more effective in terms of test set perplexity as training set size decreases.

The reason for the perplexity curves to merge as corpus size increases is obvious: the domain-specific corpus grows by including sentences less related to the biology domain. At a size of about 230M tokens, which is about half the size of the original giga-French-English corpus, the two subsampling methods are expected to have large amount of overlap, hence resulting similar perplexity values. Such convergence of performance in both subsampling methods as the corpus size increases should not be a concern because under the scenario of interest, intelligent subsampling would constitute an advantage when computational speed or storage is a limitation.

5.2 Paraphrase Examples

@@ Tables of examples @@'host' @@'channel' @@'dynamic' @@'water' or 'sodium hydroxide' or 'cholesterol' @@'activation' @@'peripheral' @@get content from

6 Conclusion

References

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proceedings of ACL.*

Regina Barzilay and Kathleen McKeown. 2001. Extracting paraphrases from a parallel corpus. In *Proceedings of ACL.*

Rahul Bhagat and Deepak Ravichandran. 2008. Large scale acquisition of paraphrases for learning surface patterns. In *Proceedings of ACL-HLT.*

Thorsten Brants and Alex Franz. 2006. Web 1T 5-gram version 1.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Josh Schroeder. 2009. Findings of the 2009 Workshop on Statistical Machine Translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 1–28, Athens, Greece, March. Association for Computational Linguistics.

Chris Callison-Burch. 2008. Syntactic constraints on paraphrases extracted from parallel corpora. In *Proceedings of EMNLP.*

Moses Charikar. 2002. Similarity estimation techniques from rounding algorithms. In *Proceedings of STOC.*

Kenneth Church and Patrick Hanks. 1991. Word association norms, mutual information and lexicography. *Computational Linguistics*, 6(1):22–29.

Hal Daumé III, Abhishek Kumar, and Avishek Saha. 2010. Frustratingly easy semi-supervised domain adaptation. In *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*, pages 53–59, Uppsala, Sweden, July. Association for Computational Linguistics.

Piotr Indyk and Rajeev Motwani. 1998. Approximate nearest neighbors: towards removing the curse of dimensionality. In *Proceedings of STOC.*

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of HLT/NAACL.*

Mirella Lapata and Frank Keller. 2005. Web-based models for natural language processing. *ACM Transactions on Speech and Language Processing*, 2(1).

Zhifei Li, Chris Callison-Burch, Chris Dyer, Sanjeev Khudanpur, Lane Schwartz, Wren Thornton, Jonathan Weese, and Omar Zaidan. 2009. Joshua: An open source toolkit for parsing-based machine translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 135–139, Athens, Greece, March. Association for Computational Linguistics.

Dekang Lin and Patrick Pantel. 2001. Discovery of inference rules for question answering. *Natural Language Engineering*, 7:343–360.

Dekang Lin, Kenneth Church, Heng Ji, Satoshi Sekine, David Yarowsky, Shane Bergsma, Kailash Patil, Emily Pitler, Rachel Lathbury, Vikram Rao, Kapil Dalwani, and Sushant Narsale. 2010. New tools for web-scale n-grams. In *Proceedings of LREC.*

Robert C. Moore and William Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 220–224, Uppsala, Sweden, July. Association for Computational Linguistics.

Bo Pang, Kevin Knight, and Daniel Marcu. 2003. Syntax-based alignment of multiple translations: Extracting paraphrases and generating new sentences. In *Proceedings of HLT/NAACL.*

Marius Pasca and Peter Dienes. 2005. Aligning needles in a haystack: Paraphrase acquisition across the web. In *Proceedings of IJCNLP*, pages 119–130.

Chris Quirk, Chris Brockett, and William Dolan. 2004. Monolingual machine translation for paraphrase generation. In *Proceedings of EMNLP*, pages 142–149.

S. Rajasekaran, S. Balla, C. Huang, V. Thapar, M. Gryk, M. Maciejewski, and M. Schiller. 2005. Selecting articles from the language model training corpus. In *In Proc. ICASSP*, pages 1695–1698.

Deepak Ravichandran, Patrick Pantel, and Eduard Hovy. 2005. Randomized Algorithms and NLP: Using Locality Sensitive Hash Functions for High Speed Noun Clustering. In *Proceedings of ACL*.

Josh Schroeder. 2007. Experiments in domain adaptation for statistical machine translation. In *Prague, Czech Republic. Association for Computational Linguistics*, pages 224–227.

Matthew Snover, Nitin Madnani, Bonnie Dorr, and Richard Schwartz. 2010. TER-plus: paraphrase, semantic, and alignment enhancements to translation edit rate. *Machine Translation*, 23(2-3):117–127.

Benjamin Van Durme and Ashwin Lall. 2010. Online generation of locality sensitive hash signatures. In *Proceedings of ACL, Short Papers*.

David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. pages 189–196.

Shiqi Zhao, Cheng Niu, Ming Zhou, Ting Liu, and Sheng Li. 2008. Combining multiple resources to improve SMT-based paraphrasing model. In *Proceedings of ACL/HLT*.