Domain-Specific Paraphrase Extraction

Ellie Pavlick[1] Juri Ganitkevitch[2] Tsz Ping Chan[3] Xuchen Yao[4]
Benjamin Van Durme[2,5] Chris Callison-Burch[1]
[1]Computer and Information Science Department, University of Pennsylvania
[2]Center for Language and Speech Processing, Johns Hopkins University
[3]Bloomberg L.P., New York, NY
[4]kitt.ai*, Seattle, WA
[5]Human Language Technology Center of Excellence, Johns Hopkins University

Abstract

The validity of applying paraphrase rules depends on the domain of the text that they are being applied to. We develop a novel method for extracting domain-specific paraphrases. We adapt the bilingual pivoting paraphrase method to bias the training data to be more like our target domain of biology. Our best model results in higher precision while retaining complete recall, giving a 10% relative improvement in AUC.

1 Introduction

Many data-driven paraphrase extraction algorithms have been developed in recent years (Madnani and Dorr, 2010; Androutsopoulos and Malakasiotis, 2010). These algorithms attempt to learn paraphrase rules, where one phrase can be replaced with another phrase which has equivalent meaning in at least some context. Determining whether a paraphrase is appropriate for a specific context is a difficult problem (Bhagat and Hovy, 2013), encompassing issues of syntax (Callison-Burch, 2008), word sense (Apidianaki et al., 2014), and style (Xu et al., 2012; Pavlick and Nenkova, 2015). To date, the question of how domain effects paraphrase has been left unexplored.

Although most paraphrase extraction algorithms attempt to estimate a confidence with which a paraphrase rule might apply, these scores are not differentiated by domain, and instead correspond to the general domain represented by the model's training data. As illustrated by Table 1, paraphrases that are highly probable in the general domain (e.g. *hot* = *sexy*) can be extremely improbable in more specialized domains like biology. Dominant word senses change depending on domain: the verb *treat* is used in expressions like *treat you to dinner* in conversational domains versus *treat an infection* in biology. This domain shift changes the acceptability of its paraphrases.

	General	Biology
hot	warm, sexy, exciting	heated, warm, thermal
treat	address, handle, buy	cure, fight, kill
head	leader, boss, mind	skull, brain, cranium

Table 1: Examples of domain-sensitive paraphrases. Most paraphrase extraction techniques learn paraphrases for a mix of senses that work well in general. But in specific domains, paraphrasing should be sensitive to specialized language use.

We address the problem of customizing paraphrase models to specific target domains. We explore the following ideas:

1. We sort sentences in the training corpus based on how well they represent the target domain, and then extract paraphrases from a subsample of the most domain-like data.

2. We improve our domain-specific paraphrases by weighting each training example based on its domain score, instead of treating each example equally.

3. We dramatically improve recall while maintaining precision by combining the subsampled in-domain paraphrase scores with the general-domain paraphrase scores.

2 Background

The paraphrase extraction algorithm that we customize is the bilingual pivoting method (Bannard and Callison-Burch, 2005) that was used to create PPDB, the paraphrase database (Ganitkevitch et al., 2013). To perform the subsampling, we adapt and improve the method that Moore and Lewis (2010) originally developed for domain-specific language models in machine translation.

*Incubated by the Allen Institute for Artificial Intelligence.

Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers), pages 57–62, Beijing, China, July 26-31, 2015.

2.1 Paraphrase extraction

Paraphrases can be extracted via bilingual pivoting. Intuitively, if two English phrases e_1 and e_2 translate to the same foreign phrase f, we can assume that e_1 and e_2 have similar meaning, and thus we can "pivot" over f and extract $\langle e_1, e_2 \rangle$ as a paraphrase pair. Since many possible paraphrases are extracted in this way, and since they vary in quality (in PPDB, the verb *treat* has 1,160 potential paraphrases, including *address*, *handle*, *deal with*, *care for*, *cure him*, *'m paying*, and *'s on the house*), it is necessary to assign some measure of confidence to each paraphrase rule. Bannard and Callison-Burch (2005) defined a conditional paraphrase probability $p(e_2|e_1)$ by marginalizing over all shared foreign-language translations f:

$$p(e_2|e_1) \approx \sum_f p(e_2|f)p(f|e_1) \qquad (1)$$

where $p(e_2|f)$ and $p(f|e_1)$ are translation model probabilities estimated from the bilingual data.

Equation 1 approximates the probability with which e_1 can paraphrase as e_2, but its estimate inevitably reflects the domain and style of the bilingual training text. If e_1 is a polysemous word, the highest probabilities will be assigned to paraphrases of the most frequently occurring sense of e_1, and lower probabilities to less frequent senses. This results in inaccurate probability estimates when moving to a domain with different sense distributions compared to the training corpus.

2.2 Sorting by domain specificity

The crux of our method is to train a paraphrase model on data from the same domain as the one in which the paraphrases will be used. In practice, it is unrealistic that we will be able to find bilingual parallel corpora precompiled for each domain of interest. We instead subsample from a large bitext, biasing the sample towards the target domain.

We adapt and extend a method developed by Moore and Lewis (2010) (henceforth M-L), which builds a domain-specific sub-corpus from a large, general-domain corpus. The M-L method assigns a score to each sentence in the large corpus based on two language models, one trained on a sample of target domain text and one trained on the general domain. We want to identify sentences which are similar to our target domain and dissimilar from the general domain. M-L captures this notion using the difference in the cross-entropies according to each language model (LM). That is, for a sentence s_i, we compute

$$\sigma_i = H_{tgt}(s_i) - H_{gen}(s_i) \qquad (2)$$

where H_{tgt} is the cross-entropy under the in-domain language model and H_{gen} is the cross-entropy under the general domain LM. Cross-entropy is monotonically equivalent to LM perplexity, in which lower scores imply a better fit. Lower σ_i signifies greater domain-specificity.

3 Domain-Specific Paraphrases

To apply the M-L method to paraphrasing, we need a sample of in-domain monolingual text. This data is not directly used to extract paraphrases, but instead to train an n-gram LM for the target domain. We compute σ_i for the English side of every sentence pair in our bilingual data, using the target domain LM and the general domain LM. We sort the entire bilingual training corpus so that the closer a sentence pair is to the top of the list, the more specific it is to our target domain.

We can apply Bannard and Callison-Burch (2005)'s bilingual pivoting paraphrase extraction algorithm to this sorted bitext in several ways:

1. By choosing a threshold value for σ_i and discarding all sentence pairs that fall outside of that threshold, we can extract paraphrases from a subsampled bitext that approximates the target domain.

2. Instead of simply extracting from a subsampled corpus (where each training example is equally weighted), we can weight each training example proportional to σ_i when computing the paraphrase scores.

3. We can combine multiple paraphrase scores: one derived from the original corpus and one from the subsample. This has the advantage of producing the full set of paraphrases that can be extracted from the entire bitext.

4 Experimental Conditions

Domain data We evaluate our domain-specific paraphrasing model in the target domain of biology. Our monolingual in-domain data is a combination of text from the GENIA database (Kim et al., 2003) and text from an introductory biology textbook. Our bilingual general-domain data is the 10^9 word parallel corpus (Callison-Burch et al.,

2009), a collection of French-English parallel data covering a mix of genres from legal text (Steinberger et al., 2006) to movie subtitles (Tiedemann, 2012). We use 5-gram language models with Kneser-Ney discounting (Heafield et al., 2013).

Evaluation We measure the precision and recall of paraphrase pairs produced by each of our models by collecting human judgments of what paraphrases are acceptable in sentences drawn from the target domain and in sentences drawn from the general domain. We sample 15K sentences from our biology data, and 10K general-domain sentences from Wikipedia. We select a phrase from each sentence, and show the list of candidate paraphrases[1] to 5 human judges. Judges make a binary decision about whether each paraphrase is appropriate given the domain-specific context. We consider a paraphrase rule to be good in the domain if it is judged to be good in least one context by the majority of judges. See Supplementary Materials for a detailed description of our methodology.

Baseline We run normal paraphrase extraction over the entire 10^9 word parallel corpus (which has 828M words on the English side) without any attempt to bias it toward the target domain. We refer this system as **General**.

Subsampling After sorting the 10^9 word parallel corpus by Equation 2, we chose several threshold values for subsampling, keeping only top-ranked τ words of the bitext. We train models on for several values of τ (1.5M, 7M, 35M, and 166M words). We refer to these model as **M-L,T=**τ.

M-L Change Point We test a model where τ is set at the point where σ_i switches from negative to positive. This includes all sentences which look more like the target domain than the general. This threshold is equivalent to sampling 20M words.

Weighted Counts Instead of weighting each subsampled sentence equally, we test a novel extension of M-L in which we weight each sentence proportional to σ_i when computing $p(e_2|e_1)$.

Combined Models We combine the subsampled models with the general model, using binary logistic regression to combine the $p(e_2|e_1)$ estimate of the general model and that of the domain-specific model. We use 1,000 labeled pairs from the target domain to set the regression weights. This tuning set is disjoint from the test set.

[1]The candidates paraphrases constitute the full set of paraphrases that can be extracted from our training corpus.

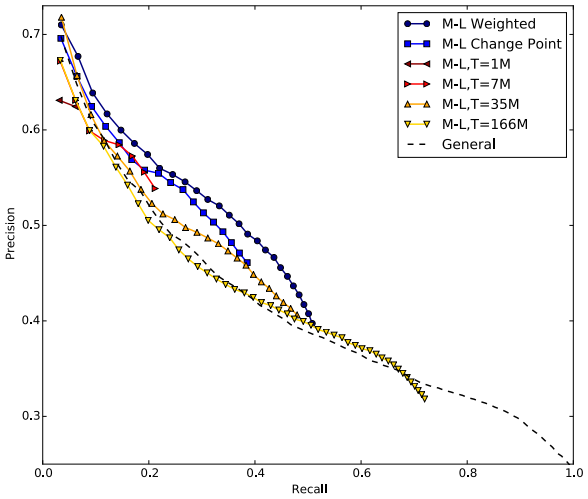


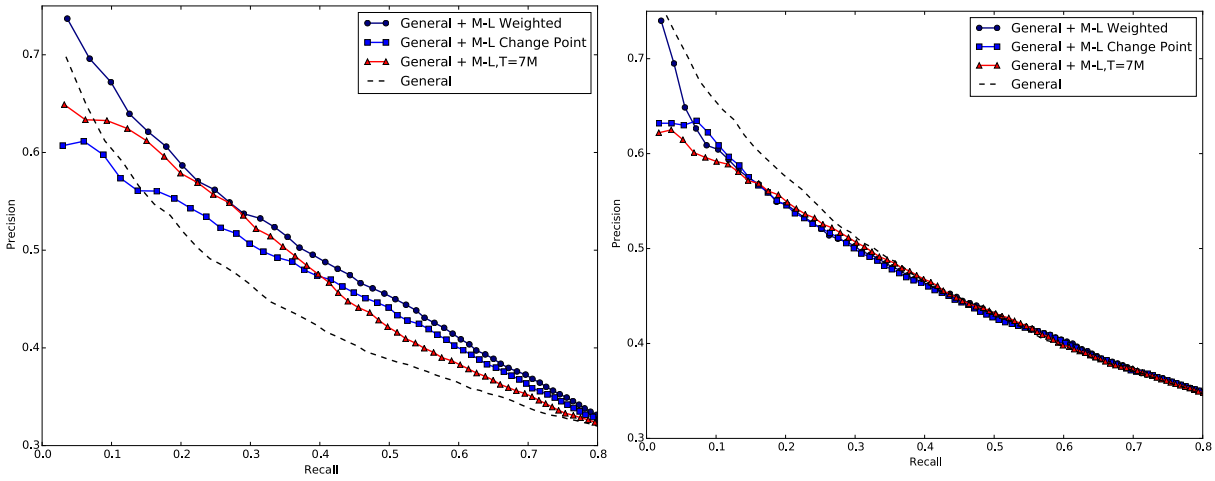
Figure 1: Precision-recall curves for paraphrase pairs extracted by models trained on data from each of the described subsampling methods. These curves are generated using the 15k manually annotated sentences in the biology domain.

5 Experimental Results

What is the effect of subsampling? Figure 1 compares the precision and recall of the different subsampling methods against the baseline of training on everything, when they are evaluated on manually labeled test paraphrases from the biology domain. All of subsampled models have a higher precision than the baseline **General** model, except for the largest of the subsampled models (which was trained on sentence pairs with 166M words - many of which are more like the general domain than the biology domain).

The subsampled models have reduced recall since many of the paraphrases that occur in the full 10^9 word bilingual training corpus do not occur in the subsamples. As we increase τ we improve recall at the expense of precision, since we are including training data that is less and less like our target domain. The highest precision model based on the vanilla M-L method is **M-L Change Point**, which sets the subsample size to include exactly those sentence pairs that look more like the target domain than the general domain.

Our novel extension of the M-L model (**M-L Weighted**) provides further improvements. Here, we weight each sentence pair in the bilingual training corpus proportional to σ_i when computing the paraphrase scores. Specifically, we weight the counting during the bilingual pivoting so that



(a) Biology domain

(b) General domain

Figure 2: Performance of models build by combining small domain-specific models trained on subsampled data with general domain models trained on all the data. Performance in the general domain are shown as a control.

rather than each occurrence counting as 1, each occurrence counts as the ratio of the sentence's cross-entropies: $\frac{H_{gen}}{H_{tgt}}$. The top-ranked sentence pairs receive an exaggerated count of 52, while the bottom ones receive a tiny factional count of 0.0068. Thus, paraphrases extracted from sentence pairs that are unlike the biology domain receive very low scores. This allows us to achieve higher recall by incorporating more training data, while also improving the precision.

What is the benefit of combining models? We have demonstrated that extracting paraphrases from subsampled data results in higher precision domain-specific paraphrases. But these models extract only a fraction of the paraphrases that are extracted by a general model trained on the full bitext, resulting in a lower recall.

We dramatically improve the recall of our domain-specific models by combining the small subsampled models with the large general-domain model. We use binary logistic regression to combine the $p(e_2|e_1)$ estimate of the general model with that of each domain-specific model. Figure 2(a) shows that we are able to extend the recall of our domain-specific models to match the recall of the full general-domain model. The precision scores remain higher for the domain-specific models. Our novel **M-L Weighted** model performs the best. Table 3 gives the area under the curve (AUC). The best combination improves AUC by more than 4 points absolute (>10 points relative) in the biology domain. Table 2 provides examples of paraphrases extracted using our domain-specific model for biology versus the baseline model.

	general / bio-spec.		general / bio-spec.
air	aerial / atmosphere	fruit	result / fruiting
balance	pay / equilibrate	heated	lively / hot
breaks	pauses / ruptures	motion	proposal / movement

Table 2: Top paraphrase under the general and the best domain-specific model, General+M-L Weighted.

	AUC	$\Delta_{absolute}$	$\Delta_{relative}$
General	39.5	–	–
Gen.+M-L,T=1	40.8	+1.3	+3.3
Gen.+M-L,T=145	40.8	+1.3	+3.3
Gen.+M-L,T=29	41.2	+1.7	+4.3
Gen.+M-L CP	41.9	+2.4	+6.1
Gen.+M-L,T=6	42.3	+2.8	+7.1
Gen.+M-L Weighted	**43.7**	**+4.2**	**+10.6**

Table 3: AUC (× 100) for each model in the biology domain from Figure 2(a).

6 Related Work

Domain-specific paraphrasing has not received previous attention, but there is relevant prior work on domain-specific machine translation (MT). We build on the Moore-Lewis method, which has been used for language models (Moore and Lewis, 2010) and translation models (Axelrod et al., 2011). Similar methods use LM perplexity to rank sentences (Gao et al., 2002; Yasuda et al., 2008), rather than the difference in cross-entropy. Within MT, Foster and Kuhn (2007) used log-linear weightings of translation probabilities to combine models trained in different domains, as we do here. Relevant to our proposed method of

fractional counting, (Madnani et al., 2007) used introduced a count-centric approach to paraphrase probability estimation. Matsoukas et al. (2009) and Foster et al. (2010) explored weighted training sentences for MT, but set weights discriminatively based on sentence-level features.

7 Conclusion

We have discussed the new problem of extracting domain-specific paraphrases. We adapt a method from machine translation to the task of learning domain-biased paraphrases from bilingual corpora. We introduce two novel extensions to this method. Our best domain-specific model dramatically improves paraphrase quality for the target domain.

Acknowledgements This research was supported by the Allen Institute for Artificial Intelligence (AI2), the Human Language Technology Center of Excellence (HLTCOE), and by gifts from the Alfred P. Sloan Foundation, Google, and Facebook. This material is based in part on research sponsored by the NSF under grant IIS-1249516 and DARPA under agreement number FA8750-13-2-0017 (the DEFT program). The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes. The views and conclusions contained in this publication are those of the authors and should not be interpreted as representing official policies or endorsements of DARPA or the U.S. Government.

We would like to thank Luke Orland for his contributions to this research, and to thank the anonymous reviewers for their thoughtful comments.

References

Ion Androutsopoulos and Prodromos Malakasiotis. 2010. A survey of paraphrasing and textual entailment methods. *JAIR*, pages 135–187.

Marianna Apidianaki, Emilia Verzeni, and Diana McCarthy. 2014. Semantic clustering of pivot paraphrases. In *LREC*.

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *EMNLP*, pages 355–362.

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *ACL*, pages 597–604.

Rahul Bhagat and Eduard Hovy. 2013. What is a paraphrase? *Computational Linguistics*, 39(3):463–472.

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*.

Chris Callison-Burch. 2008. Syntactic constraints on paraphrases extracted from parallel corpora. In *EMNLP*, pages 196–205. Association for Computational Linguistics.

George Foster and Roland Kuhn. 2007. Mixture-model adaptation for smt.

George Foster, Cyril Goutte, and Roland Kuhn. 2010. Discriminative instance weighting for domain adaptation in statistical machine translation. In *EMNLP*, pages 451–459.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. PPDB: The paraphrase database. In *NAACL-HLT*, pages 758–764, Atlanta, Georgia, June.

Jianfeng Gao, Joshua Goodman, Mingjing Li, and Kai-Fu Lee. 2002. Toward a unified approach to statistical language modeling for chinese. *ACM Transactions on Asian Language Information Processing (TALIP)*, 1(1):3–33.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H Clark, and Philipp Koehn. 2013. Scalable modified kneser-ney language model estimation. In *ACL*.

J-D Kim, Tomoko Ohta, Yuka Tateisi, and Junichi Tsujii. 2003. Genia corpusa semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl 1):i180–i182.

Nitin Madnani and Bonnie J. Dorr. 2010. Generating phrasal and sentential paraphrases: A survey of data-driven methods. *Computational Linguistics*, 36.

Nitin Madnani, Necip Fazil Ayan, Philip Resnik, and Bonnie Dorr. 2007. Using paraphrases for parameter tuning in statistical machine translation. In *Proceedings of the Workshop on Machine Translation*.

Spyros Matsoukas, Antti-Veikko I Rosti, and Bing Zhang. 2009. Discriminative corpus weight estimation for machine translation. In *EMNLP*, pages 708–717.

Robert C Moore and William Lewis. 2010. Intelligent selection of language model training data. In *ACL*, pages 220–224.

Ellie Pavlick and Ani Nenkova. 2015. Inducing lexical style properties for paraphrase and genre differentiation. In *NAACL*.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaz Erjavec, Dan Tufis, and Dániel Varga. 2006. The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages. *arXiv preprint*.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *LREC*, pages 2214–2218.

Wei Xu, Alan Ritter, Bill Dolan, Ralph Grishman, and Colin Cherry. 2012. Paraphrasing for style. In *COLING*, pages 2899–2914.

Keiji Yasuda, Ruiqiang Zhang, Hirofumi Yamamoto, and Eiichiro Sumita. 2008. Method of selecting training data to build a compact and efficient translation model. In *IJCNLP*, pages 655–660.