Chapter 1

Introduction

Statistical machine translation (SMT) systems are trained using a large collection of translated sentence pairs known as a parallel corpus. Common sources of parallel data include parliament proceedings, books, and news articles. For some language pairs, we have large amounts of this data. For example, the Canadian Hansards are parliamentary proceedings that give us millions of words of French/English parallel data (Germann, 2001a). Similarly, the proceedings of European Union parliament are a source of parallel data for all of the languages of its member states (Koehn, 2005). Outside of these government sources, we also have large collections of parallel data from news agencies for some language pairs, such as Chinese/English (Ma, 2005). However, for most language pairs, we have little to no data available. In addition, even when parallel data is available, it often does not match the domain of the data you wish to translate, which hurts translation quality (Munteanu and Marcu, 2005).

[This section needs some examples to make it crystal clear what you're talking about, even for a non-NLP person. Show an example in news or government, and one of your other domains (wikipedia, travel, etc.)]

The creation of new parallel corpora can be expensive, especially when bilingual speakers are rare for the language pair of interest. Germann (2001b) investigated the costs of collecting enough data to build Tamil/English SMT system. They found that professionally translated data would cost $0.36 per word. Germann (2001b) and others (Zaidan and Callison-Burch, 2011) were able to reduce the cost of creating parallel corpora by looking to non-professional translators, but the cost is still around $0.10 per word. In order to acquire more parallel data without this cost, researchers have looked to multilingual corpora which share some content across languages, but are not directly translated. Such corpora are referred to as comparable corpora, and examples include multilingual news feeds (Munteanu and Marcu, 2005), Wikipedia articles (Adafre and de Rijke, 2006; Smith et al., 2010), and the Web (Resnik, 1999; Nie et al., 1999; Chen and Nie, 2000).

Comparable corpora is a broad term—Fung and Cheung (2004a) give a more fine-grained categorization of multilingual corpora:

1. Parallel corpus: A sentence-aligned corpus containing bilingual translations of the same document. (Curated parallel corpora)

2. Noisy parallel corpus: A corpus containing non-aligned sentences that are nevertheless mostly bilingual translations of the same document. (Hansards, Eu-

roparl, most “parallel” corpora)

3. Comparable corpus: A corpus of non-aligned and non-translated documents which are topic-aligned. (Wikipedia)

4. Quasi-comparable corpus: A multilingual corpus which is not sentence-aligned, translated, or topic-aligned. (the Web, multilingual news feeds) [I don’t even know what this means. Does it serve your larger point to mention this?] **[The purpose of these categories is just to show that very different corpora are called comparable, and they require very different methods to mine for parallel data. I added examples to each category, but I’m not sure if that helps with the confusion. –JS]**

As comparable corpora vary greatly in their structure, different methods for finding parallel sentences are used in each.

Even corpora which are generally considered as parallel require some amount of processing to find parallel sentences. A translator may chose to translate a compound sentence as two sentences, or vice-versa, so naively assuming that sentences are aligned in order will not work. Also, there may be large insertions or deletions of sentences even in curated sources of parallel data, such as the Canadian Hansards (Gale and Church, 1993; Chen, 1993). Sentence-aligning these corpora does not require existing parallel data or a bilingual dictionary for the language pair of interest. Instead, the structure of the documents and the lengths of the sentences are used to determine the sentence alignment. For comparable corpora which are topic-aligned but not directly translated, lexical information must be used to determine which sentence pairs should be aligned (Munteanu and Marcu, 2005). When comparable corpora are not topic-aligned, other signals are exploited to find plausible document alignments (Resnik and Smith, 2003).

We will examine a representative set of comparable corpora: the Web, Twitter, and Wikipedia; describe the different signals used to identify parallel data, and demonstrate how extracted parallel data from these corpora improve SMT performance across several language pairs and domains. First, we scale up previous Web mining methods (Resnik and Smith, 2003) to several terabytes of data. We also present a novel mining approach for Twitter, making use of metadata unique to the microblogging medium. Finally, we introduce a new sentence alignment model for mining parallel data from Wikipedia which takes advantage of its fine-grained topic alignment.[I am not sure about this particular order. Let's discuss soon.]

1.1 What counts as parallel?

This work is centered around finding parallel data—bilingual sentence pairs which convey the same meaning. Unfortunately, it is extremely difficult, if not impossible, to determine whether or not two sentences in different languages have the same meaning. One language may contain gender markings that the other does not, or the connotation of a word may be difficult to express in another language. Examples of this problem are explored in depth by Kay (1997). Even ignoring the cross-lingual issues, comparing the meaning of two sentences in the same language is still quite difficult—SMT evaluation metrics (Papineni et al., 2002; Banerjee and Lavie, 2005; Snover et al., 2006) must address this problem. [I like this, but note that this stands somewhat in opposition to the Fung paper you talk about above. This gives a very functional description of parallelism: it's anything that makes the BLEU score go up. Of course your evaluations might be somewhat insensitive to errors in detecting parallelism—we don't know if they care more about precision or recall. I think you need to address this, probably empirically.] **[I don't understand how this is in conflict with the Fung paper. I only quote that to talk about different types of comparable corpora. –JS]**

When evaluating methods for finding parallel data, we can either measure intrinsic or extrinsic performance. Intrinsic evaluation directly measures the quantity and quality of parallel data we extract, while extrinsic evaluation is only concerned with how the new parallel data improves SMT performance. In order to perform intrinsic evaluation, we need some criteria for determining whether or not a bilingual sentence pair is parallel. This is easy if we use parallel data, but it is preferable to evaluate our methods on the same corpora that we are extracting data from. When designing the criteria for judging parallel sentences, we focus on our extrinsic goal: improving

SMT performance. If a sentence pair is likely to improve performance when added to our SMT system's training data, we would like to extract it. The details of our annotation criteria can be found in Chapter 3, but in all cases they are motivated by SMT performance. To understand what will influence performance, we need to understand modern SMT systems.

1.2 Statistical Machine Translation

While machine translation has been around in some form for many decades (Locke and Booth, 1955), statistical machine translation began with the work of Brown et al. (1988, 1990, 1993). SMT systems have evolved since then, most notably moving from word-based systems to phrase-based (Koehn et al., 2003). Several newer systems have been developed, focusing mostly on incorporating syntax into the translation model (Chiang et al., 2005; Quirk et al., 2005; Liu et al., 2006; Galley et al., 2006). These systems all share some key characteristics in how they use parallel data:

1. A large collection of parallel sentences are used as training data.

2. For each parallel sentence, word-to-word correspondences are found. This step is called word-alignment, and it is usually done with unsupervised methods (Brown et al., 1993; Vogel et al., 1996).

3. Pairs of phrases, or other multi-word units, are extracted from the word-aligned sentence pairs to form a translation model.[Example.]

4. A language model is created from large amounts of monolingual data in the "target" language (the language which text is translated into). This includes the target side of the parallel training data.

There are additional details in each model, but the main effects of adding new parallel data are additional inputs to the translation and language models. [The language model argument is weaker here.]

1.3 Evaluation Pipeline

Our evaluation setup is identical across chapters—we start with initial data that includes some standard parallel and monolingual corpora commonly used for translation. We also have extracted parallel data that we find in a comparable corpus. Table 1.3 describes how we use this data to measure SMT improvements:

In both the baseline and experimental conditions, we include the target side of the extracted parallel sentences in the monolingual training data. We do this to ensure that any increase in performance is coming from the parallel data. It would be simple to add monolingual text from a comparable corpus to an SMT system.[This is probably also a good baseline.]

In all experiments, the BLEU metric (Papineni et al., 2002) is used to evaluate SMT performance. The BLEU metric combines n-gram precision (the percentage of n-grams in the hypothesis translation which are found in the reference) with a brevity

	Parallel	Monolingual
Baseline	Initial	Initial + Extracted
Experimental	Initial + Extracted	Initial + Extracted

Table 1.1: Parallel and monolingual data used in our SMT experiments.

[This table seems a little suspect. Why would you use initial+extracted for the baseline results? Realistically, you could have a baseline built from a large monolingual text of any kind (Maybe even in-domain). The line you list as baseline here makes sense to tease apart whether the differences come from TM or LM (that's good), but should probably include an additional baseline.] **[Another baseline without the extracted target side in the language model wouldn't hurt, but I don't think it's crucial and it's not available for the Wikipedia experiments. Everything I've done so far has used this setup. It might be something I include in some experiments. –JS]** penalty. The initial data, test sets, and other details vary by experiment.

1.4 Sentence Alignment

[I agree that this probably does not fit here. The question is: what does? I think you might want to say something about prior art here, or otherwise fit in relevant parts of your lit review. Logically, a reader that has arrived at this point in your dissertation should understand the problem you're trying to solve at a high level, and might wonder what other approaches have been taken.] In this section, we will describe our task and notation. We will view both parallel corpora alignment and the extraction of parallel sentences from comparable corpora as an alignment task. In either type of alignment we are given a set of bilingual document pairs in source and target languages. When performing parallel corpora alignment, these document pairs will correspond to each other very strongly, while in the case of comparable corpora, some these document pairs may contain no parallel sentences. Munteanu and Marcu (2005) take their document pairs from news stories published at roughly the same time, while Adafre and de Rijke (2006); Smith et al. (2010) use entries from Wikipedia that are on the same topic (Figure 1.1 gives and example). The task of finding comparable document pairs is not addressed in this work.

Each document pair contains a sequence of source sentences (denoted by $\mathbf{S}$) and target sentences (denoted by $\mathbf{T}$). Individual source and target sentences are referred


Antipartícula

A cada una de las partículas de la naturaleza le corresponde una **antipartícula** que posee la misma masa, el mismo espín, pero distinta carga eléctrica. Algunas partículas son idénticas a su antipartícula, como por ejemplo el fotón, que no tiene carga. Pero no todas las partículas de carga neutra son idénticas a su antipartícula.

Antiparticle

From Wikipedia, the free encyclopedia

Corresponding to most kinds of particles, there is an associated **antiparticle** with the same mass and opposite electric charge. For example, the antiparticle of the electron is the positively charged antielectron, or positron, which is produced naturally in certain types of radioactive decay.

Figure 1.1: An example of a Spanish/English document pair from Wikipedia.[I think you want this example to appear in your comparable corpus section.]

to by S and T respectively. Similarly, we refer to the words within source and target sentences with the lowercase s and t. We borrow the notation of (Och and Ney, 2003) for describing alignments between sentences as subsets of the Cartesian product of sentence positions. Sentence alignments are referred to with the uppercase A, and word alignments with the lowercase a.

The goal of sentence alignment is to identify which sentence pairs in the bilingual document pairs are parallel. We view this as a retrieval task for parallel sentence pairs, and so when annotated sentence alignments are present, we can compute precision, recall, and F-measure.

Bibliography

Sadaf Abdul-Rauf and Holger Schwenk. On the use of comparable corpora to improve smt performance. In Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, EACL '09, pages 16–23, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=1609067.1609068.

Sisay F. Adafre and Maarten de Rijke. Finding Similar Sentences Across Multiple Languages in Wikipedia. In EACL 2006 Workshop on New Text, Wikis and Blogs and Other Dynamic Text Sources, 2006.

Vamshi Ambati and Stephan Vogel. Can crowds build parallel corpora for machine translation systems? In Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, CSLDAMT '10, pages 62–65, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=1866696.1866706.

Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pages 65–72, 2005.

Phil Blunsom and Trevor Cohn. Discriminative word alignment with conditional random fields. In Proceedings of ACL, 2006.

Peter Brown, John Cocke, S Della Pietra, V Della Pietra, Frederick Jelinek, Robert Mercer, and Paul Roossin. A statistical approach to language translation. In Proceedings of the 12th conference on Computational linguistics-Volume 1, pages 71–76. Association for Computational Linguistics, 1988.

Peter F Brown, John Cocke, Stephen A Della Pietra, Vincent J Della Pietra, Fredrick Jelinek, John D Lafferty, Robert L Mercer, and Paul S Roossin. A statistical approach to machine translation. Computational linguistics, 16(2):79–85, 1990.

Peter F. Brown, Jennifer C. Lai, and Robert L. Mercer. Aligning sentences in parallel corpora. In Proceedings of the 29th annual meeting on Association for Computational Linguistics, ACL '91, pages 169–176, Stroudsburg, PA, USA, 1991. Association for Computational Linguistics. doi: 10.3115/981344.981366. URL `http://dx.doi.org/10.3115/981344.981366`.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. The mathematics of statistical machine translation: Parameter estimation. Comput. Linguist., 1993.

Chris Callison-Burch. Fast, cheap, and creative: evaluating translation quality using amazon's mechanical turk. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1, EMNLP '09, pages 286–295, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-59-6. URL http://dl.acm.org/citation.cfm?id=1699510.1699548.

Jiang Chen and Jian-Yun Nie. Parallel web text mining for cross-language ir. In IN IN PROC. OF RIAO, pages 62–77, 2000.

Stanley F. Chen. Aligning sentences in bilingual corpora using lexical information. In Proceedings of the 31st annual meeting on Association for Computational Linguistics, ACL '93, pages 9–16, Stroudsburg, PA, USA, 1993. Association for Computational Linguistics. doi: http://dx.doi.org/10.3115/981574.981576. URL http://dx.doi.org/10.3115/981574.981576.

David Chiang, Adam Lopez, Nitin Madnani, Christof Monz, Philip Resnik, and Michael Subotin. The hiero machine translation system: Extensions, evaluation, and analysis. In Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, pages 779–786. Association for Computational Linguistics, 2005.

Kenneth Ward Church. Char align: a program for aligning parallel texts at the character level. In Proceedings of the 31st annual meeting on Association for

Computational Linguistics, ACL '93, pages 1–8, Stroudsburg, PA, USA, 1993. Association for Computational Linguistics. doi: 10.3115/981574.981575. URL http://dx.doi.org/10.3115/981574.981575.

I. Dagan, K. W Church, and W. A Gale. Robust bilingual word alignment for machine aided translation. In Proceedings of the Workshop on Very Large Corpora: Academic and Industrial Perspectives, pages 1–8, 1993.

J. Dean and S. Ghemawat. Mapreduce: simplified data processing on large clusters. In Proceedings of the 6th conference on Symposium on Opearting Systems Design & Implementation-Volume 6, pages 10–10. USENIX Association, 2004.

A.P. Dempster, N.M. Laird, D.B. Rubin, et al. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B (Methodological), 39(1):1–38, 1977.

Andreas Eisele and Yu Chen. Multiun: A multilingual corpus from united nation documents. In Daniel Tapias, Mike Rosner, Stelios Piperidis, Jan Odjik, Joseph Mariani, Bente Maegaard, Khalid Choukri, and Nicoletta Calzolari (Conference Chair), editors, Proceedings of the Seventh conference on International Language Resources and Evaluation, pages 2868–2872. European Language Resources Association (ELRA), 5 2010.

Jason Eisner. Parameter estimation for probabilistic finite-state transducers. In Proceedings of the 40th Annual Meeting on Association for Computational

Linguistics, ACL '02, pages 1–8, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073085. URL http://dx.doi.org/10.3115/1073083.1073085.

Pascale Fung and Percy Cheung. Multi-level bootstrapping for extracting parallel sentences from a quasi-comparable corpus. In Proceedings of the 20th international conference on Computational Linguistics, COLING '04, Stroudsburg, PA, USA, 2004a. Association for Computational Linguistics. doi: http://dx.doi.org/10.3115/1220355.1220506. URL http://dx.doi.org/10.3115/1220355.1220506.

Pascale Fung and Percy Cheung. Mining Very-Non-Parallel Corpora: Parallel Sentence and Lexicon Extraction via Bootstrapping and EM. In EMNLP 04, 2004b.

Pascale Fung and Kenneth Ward Church. K-vec: a new approach for aligning parallel texts. In Proceedings of the 15th conference on Computational linguistics - Volume 2, COLING '94, pages 1096–1102, Stroudsburg, PA, USA, 1994. Association for Computational Linguistics. doi: http://dx.doi.org/10.3115/991250.991328. URL http://dx.doi.org/10.3115/991250.991328.

William A. Gale and Kenneth W. Church. Identifying word correspondence in parallel texts. In Proceedings of the workshop on Speech and Natural Language, HLT '91, pages 152–157, Stroudsburg, PA, USA, 1991. Association for Computational Linguistics. doi: 10.3115/112405.112428. URL http://dx.doi.org/10.3115/112405.112428.

William A. Gale and Kenneth W. Church. A program for aligning sentences in bilingual corpora. Comput. Linguist., 19:75–102, March 1993. ISSN 0891-2017. URL http://dl.acm.org/citation.cfm?id=972450.972455.

Michel Galley, Jonathan Graehl, Kevin Knight, Daniel Marcu, Steve DeNeefe, Wei Wang, and Ignacio Thayer. Scalable inference and training of context-rich syntactic translation models. In Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, pages 961–968. Association for Computational Linguistics, 2006.

Ulrich Germann. Aligned hansards of the 36th parliament of canada. Natural Language Group of the USC Information Sciences Institute, 2001a.

Ulrich Germann. Building a statistical machine translation system from scratch: how much bang for the buck can we expect? In Proceedings of the workshop on Data-driven methods in machine translation-Volume 14, pages 1–8. Association for Computational Linguistics, 2001b.

Fei Huang, Ying Zhang, and Stephan Vogel. Mining key phrase translations from web corpora. In Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, HLT '05, pages 483–490, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics. doi: 10.3115/1220575.1220636. URL http://dx.doi.org/10.3115/1220575.1220636.

Roy Bar-Haim Ido, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor. The second pascal recognising textual entailment challenge, 2006.

Martin Kay. The proper place of men and machines in language translation. Machine Translation, 12(1-2):3–23, 1997.

Martin Kay and Martin Röscheisen. Text-translation alignment. Comput. Linguist., 19:121–142, March 1993. ISSN 0891-2017. URL `http://dl.acm.org/citation.cfm?id=972450.972457`.

P. Koehn. Europarl: A parallel corpus for statistical machine translation. In MT summit, volume 5, 2005.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. Statistical phrase-based translation. In hltnaacl, pages 127–133, Edmonton, Canada, May 2003. URL `http://people.csail.mit.edu/people/koehn/publications/phrase2003.pdf`.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris ohis has led researchers to develop methods for finding Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. Moses: open source toolkit for statistical machine translation. In Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, ACL '07, pages 177–180, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=1557769.1557821.

J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of the 18th International Conference on Machine Learning, pages 282–289, 2001.

Yang Liu, Qun Liu, and Shouxun Lin. Tree-to-string alignment template for statistical machine translation. In Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, pages 609–616. Association for Computational Linguistics, 2006.

William Nash Locke and Andrew Donald Booth. Machine translation of languages: fourteen essays. Published jointly by Technology Press of the Massachusetts Institute of Technology and Wiley, New York, 1955.

Edward Loper and Steven Bird. Nltk: the natural language toolkit. In Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics - Volume 1, ETMTNLP '02, pages 63–70, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics. doi: 10.3115/1118108.1118117. URL http://dx.doi.org/10.3115/1118108.1118117.

Xiaoyi Ma. Chinese english news magazine parallel text. LDC2005T10, 2005.

Robert Moore. Fast and Accurate Sentence Alignment of Bilingual Corpora. In Stephen Richardson, editor, Machine Translation: From Research to Real Users, volume 2499 of Lecture Notes in Computer Science, pages 135–144. Springer Berlin / Heidelberg, 2002. ISBN 978-3-540-44282-0.

Dragos Stefan Munteanu and Daniel Marcu. Improving Machine Translation Performance by Exploiting Non-Parallel Corpora. Comput. Linguist., 31:477–504, December 2005. ISSN 0891-2017. doi: http://dx.doi.org/10.1162/089120105775299168. URL `http://dx.doi.org/10.1162/089120105775299168`.

Dragos Stefan Munteanu, Alexander Fraser, and Daniel Marcu. Improved Machine Translation Performance via Parallel Sentence Extraction from Comparable Corpora. In NAACL, pages 265–272, 2004.

D.S. Munteanu and D. Marcu. Extracting parallel sub-sentential fragments from non-parallel corpora. In ANNUAL MEETING-ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, volume 44, page 81, 2006.

Jian-Yun Nie, Michel Simard, Pierre Isabelle, and Richard Durand. Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the web. In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '99, pages 74–81, New York, NY, USA, 1999. ACM. ISBN 1-58113-096-1. doi: 10.1145/312624.312656. URL `http://doi.acm.org/10.1145/312624.312656`.

Franz Josef Och. Minimum error rate training in statistical machine translation. In acl, pages 160–167, Sapporo, Japan, 2003. URL http://acl.ldc.upenn.edu/P/P03/P03-1021.pdf.

Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. Comput. Linguist., 29:19–51, March 2003. ISSN 0891-2017. doi: http://dx.doi.org/10.1162/089120103321337421. URL http://dx.doi.org/10.1162/089120103321337421.

Jose Oncina and Marc Sebban. Learning stochastic edit distance: Application in handwritten character recognition. Pattern Recogn., 39:1575–1587, September 2006. ISSN 0031-3203. doi: 10.1016/j.patcog.2006.03.011. URL http://dl.acm.org/citation.cfm?id=1220973.1221331.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In acl, pages 311–318, Philadelpha, Pennsylvania, USA, 2002.

Matt Post, Chris Callison-Burch, and Miles Osborne. Constructing parallel corpora for six indian languages via crowdsourcing. WMT '10, 2012.

Chris Quirk, Arul Menezes, and Colin Cherry. Dependency treelet translation: Syntactically informed phrasal smt. In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, pages 271–279. Association for Computational Linguistics, 2005.

Chris Quirk, Raghavendra Udupa, and Arul Menezes. Generative models of noisy translations with applications to parallel fragment extraction. In EAMT 2007, 2007.

P. Resnik and N. A Smith. The web as a parallel corpus. Computational Linguistics, 29(3):349–380, 2003.

Philip Resnik. Mining the web for bilingual text. In Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, ACL '99, pages 527–534, Stroudsburg, PA, USA, 1999. Association for Computational Linguistics. ISBN 1-55860-609-3. doi: 10.3115/1034678.1034757. URL http://dx.doi.org/10.3115/1034678.1034757.

Jason Riesa and Daniel Marcu. Automatic parallel fragment extraction from noisy data. In Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 538–542, Montréal, Canada, June 2012. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/N12-1061.

Eric Sven Ristad and Peter N. Yianilos. Learning String-Edit Distance. IEEE Trans. Pattern Anal. Mach. Intell., 20:522–532, May 1998. ISSN 0162-8828. doi: 10.1109/34.682181. URL http://dl.acm.org/citation.cfm?id=279270.279279.

Lei Shi, Cheng Niu, Ming Zhou, and Jianfeng Gao. A dom tree alignment model for mining parallel data from the web. In Proceedings of the 21st International

Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, ACL-44, pages 489–496, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics. doi: 10.3115/1220175.1220237. URL http://dx.doi.org/10.3115/1220175.1220237.

Michel Simard, George F. Foster, and Pierre Isabelle. Using cognates to align sentences in bilingual corpora. In Proceedings of the 1993 conference of the Centre for Advanced Studies on Collaborative research: distributed computing - Volume 2, CASCON '93, pages 1071–1082. IBM Press, 1993. URL http://dl.acm.org/citation.cfm?id=962367.962411.

Jason R. Smith, Chris Quirk, and Kristina Toutanova. Extracting Parallel Sentences from Comparable Corpora using Document Level Alignment. In NAACL 2010, 2010.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. In Proceedings of association for machine translation in the Americas, pages 223–231, 2006.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '08, pages 254–263, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=1613715.1613751.

Jörg Tiedemann. News from OPUS - A collection of multilingual parallel corpora with tools and interfaces. In N. Nicolov, K. Bontcheva, G. Angelova, and R. Mitkov, editors, Recent Advances in Natural Language Processing, volume V, pages 237–248. John Benjamins, Amsterdam/Philadelphia, Borovets, Bulgaria, 2009. ISBN 978 90 272 4825 1.

C. Tillmann. A Beam-Search extraction algorithm for comparable data. In Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, pages 225–228, 2009.

C. Tillmann and J. Xu. A simple sentence-level extraction algorithm for comparable data. In Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers, pages 93–96, 2009.

Ferhan Ture and Jimmy Lin. Why not grab a free lunch? mining large corpora for parallel sentences to improve translation modeling. In Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 626–630, Montréal, Canada, June 2012. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/N12-1079.

Ferhan Ture, Tamer Elsayed, and Jimmy Lin. No free lunch: brute force vs. locality-sensitive hashing for cross-lingual pairwise similarity. In Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, SIGIR '11, pages 943–952, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0757-4. doi: 10.1145/2009916.2010042. URL http://doi.acm.org/10.1145/2009916.2010042.

Jakob Uszkoreit, Jay M. Ponte, Ashok C. Popat, and Moshe Dubiner. Large scale parallel document mining for machine translation. In Proceedings of the 23rd International Conference on Computational Linguistics, COLING '10, pages 1101–1109, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=1873781.1873905.

Ashish Venugopal, Jakob Uszkoreit, David Talbot, Franz J. Och, and Juri Ganitkevitch. Watermarking the outputs of structured prediction with an application in statistical machine translation. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '11, pages 1363–1372, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. ISBN 978-1-937284-11-4. URL http://dl.acm.org/citation.cfm?id=2145432.2145576.

S. Vogel, H. Ney, and C. Tillmann. HMM-based word alignment in statistical translation. In Proceedings of the 16th conference on Computational linguistics-Volume 2, pages 836–841, 1996.

Wikipedia. Wikipedia, the free encyclopedia, 2004. URL `\url{http://en.wikipedia.org/}`. [Online; accessed 01-June-2009].

Dekai Wu and Pascale Fung. Inversion Transduction Grammar Constraints for Mining Parallel Sentences from Quasi-Comparable Corpora. In Robert Dale, Kam-Fai Wong, Jian Su, and Oi Kwong, editors, Natural Language Processing IJCNLP 2005, volume 3651 of Lecture Notes in Computer Science, pages 257–268. Springer Berlin / Heidelberg, 2005. ISBN 978-3-540-29172-5.

Omar F. Zaidan and Chris Callison-Burch. Crowdsourcing translation: professional quality from non-professionals. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1, HLT '11, pages 1220–1229, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. ISBN 978-1-932432-87-9. URL http://dl.acm.org/citation.cfm?id=2002472.2002626.

B. Zhao and S. Vogel. Adaptive parallel sentences mining from web bilingual news collection. In Proceedings of the 2002 IEEE International Conference on Data Mining, page 745. IEEE Computer Society, 2002.