Chapter 2

Related Work

Research in SMT began as large parallel corpora became available (Brown et al., 1988, 1990, 1993). These corpora include the Canadian Hansards (French-English parliament proceedings) and the Hong Kong Laws Corpus, among many others. While these corpora were parallel in the sense that they were created by directly translating text in one language, they were not sentence aligned. Noise in the form of missing data or sentences without a $1:1$ correspondence made alignment a non-trivial problem. This lead to the development of several approaches for aligning parallel corpora in the early 1990s. Since the problem of aligning noisy-parallel corpora is closely related to finding parallel sentences in comparable corpora, we will give an overview of these approaches.

2.1 Parallel Corpus Alignment

Perhaps the most well known work on parallel corpus alignment is Gale and Church (1991, 1993). The authors described a sentence alignment method based on dynamic programming which used only sentence length to determine whether or not two sentences were parallel. This method is widely applicable since it assumes almost no linguistic knowledge.[1] Despite this, it achieved very high accuracy on a corpus of economic reports from the Union Bank of Switzerland in English, French and German. Brown et al. (1991) had a similar approach, using only sentence lengths to align parallel corpora, but they measured length in words rather than characters.

Even when there is no bilingual lexicon available for a language pair, if the source and target languages are similar enough it may be possible to use the surface similarity of words to infer cognates. Simard et al. (1993) made use of this by replacing the length based alignment scoring of Gale and Church (1993) with a cognate based scoring method using a simple method for identifying cognates. Church (1993) made use of cognates with a radically different approach: creating a dotplot of character n-gram matches weighted by inverse frequency, and then finding an alignment which best matches the dots. While this cognate based approach was intended to work for similar languages, the authors noted that even in language pairs like Japanese-English, matches can be found on technical terms and markup.

[1] The only bit of information about the language pair required is a ratio of sentence lengths in characters.

The sentence alignment approach of Kay and Röscheisen (1993) also used little linguistic knowledge, though they build a bilingual dictionary from the parallel text to facilitate alignment. Beginning with an initial set of sentence alignments, they iteratively update the bilingual dictionary and the sentence alignments in a manner similar to Viterbi EM, though no explicit probability model is given. Chen (1993) had a similar approach, except he incorporated the learning of both sentence and word alignments into a probabilistic model. While this is similar to our work in that there is a generative story of document pairs used to infer sentence alignments, Chen (1993) used a joint probability distribution of source/target sentence pairs which must be approximated for efficient inference, and several choices are made in the inference strategy which assume a strongly monotonic sentence alignment. Stochastic Viterbi EM is used to find the best sentence alignment.

As an alternative method for creating a bilingual dictionary, Fung and Church (1994) built a vector for each source/target word representing how it is distributed in the parallel corpus. The intuition was that since the alignment between the source and target data was strongly monotonic, so words that appear in the same relative positions in the source/target corpora are likely to be translations of one another.

Moore (2002) builds off of the length based alignment approach of Gale and Church (1993) by adding a bootstrapping step after the initial alignment. First, a length based sentence alignment is done on the parallel corpus. Then, the sentences found to be parallel are used to train a word alignment model (IBM Model 1), and the sentence

alignment dynamic program is repeated using the word alignment scores in addition to length based scores. This bootstrapping approach is popular in work on mining noisy parallel/comparable corpora (see Section 2.2).

2.2 Comparable Corpus Mining

In addition to aligning parallel texts, there has also been a considerable amount of work done on finding parallel sentence pairs in comparable corpora. A comparable corpus is a multilingual collection of documents which may contain parallel sentences, but is not completely parallel. This broad definition includes both weakly aligned data such as timestamped multilingual news feeds, and Wikipedia articles linked at the document level. Depending on the type of comparable corpus, different methods may be more or less effective for finding parallel sentences. We will split our review of comparable corpora mining methods into two categories. In Section 2.2.1, we will examine methods used on closely aligned comparable corpora, and in Section 2.2.2 we will review work on extracting parallel sentences from less related multilingual documents.

2.2.1 Noisy Parallel Corpora

The first category of work on comparable corpora mining that we will review is on noisy parallel data. While even corpora called "parallel" contain some noise, we

are refering to corpora which the methods in Section 2.1 would fail on.

Similar to the dynamic programming approaches explored in Section 2.1, Zhao and Vogel (2002) used a dynamic programming strategy for aligning parallel sentences in a document pair. They create a probabilistic model of a comparable document pair $P(S, T, A)$ and choose an alignment to maximize the probability of the observed source and target documents. To estimate the probability of two sentences being aligned, they used and IBM-style word alignment models (Model 3, specifically) which were estimated on existing parallel data. Zhao and Vogel (2002) also describes a bootstraping approach where high confidence sentence alignments are added to the training data for the word alignment model, and then sentence alignments are recomputed. Much of the work on noisy parallel/comparable corpora mining used this technique (Fung and Cheung, 2004a,b; Wu and Fung, 2005; Munteanu and Marcu, 2005).

2.2.2 Comparable Corpora

In comparable corpora such as bilingual news feeds or websites, the document alignment is often not given.[2] First, we will review methods for finding comparable document pairs in a comparable corpus, and then methods for identifying parallel sentence pairs within these documents.

[2]A notable exception to this is Wikipedia

2.2.2.1 Finding Comparable Document Pairs

The Gigaword corpus contains news feeds in multiple languages, and is annotated with the date of publication. Since these news articles are potentially on the same topic, there are potentially parallel sentence pairs in these articles. Munteanu et al. (2004); Munteanu and Marcu (2005); Fung and Cheung (2004a,b) make use of this information to find comparable document pairs. The basic strategy is to first consider all bilingual article pairs published within a time window to be potentially comparable. Then, documents in one language are projected through a bilingual dictionary, and bag-of-words based document similarity measures are used to prune this large set of document pairs. This requires either existing parallel data or at least a bilingual dictionary. Document pairs that pass through these filter are then mined for parallel sentences.

Multilingual websites are another potential source for comparable or parallel document pairs. STRAND (Resnik and Smith, 2003) used some heuristics for identifying links between versions of the same website in different languages. This provides a candidate set of document pairs, which are further filtered by looking at their HTML structure. Each website is converted into a list of start tags, end tags, and "chunks" (text within a tag), and these lists are aligned using standard dynamic programming techniques. This alignment is not only used to determine whether a pair of websites is comparable, but it also gives an alignment of text chunks which greatly narrows down the space of possible sentence alignments

A drastically different approach for finding parallel web pages is given by Uszkoreit et al. (2010). Using a existing language identification and translation systems, they identify the language of all webpages and translate the non-English ones into English. Since all documents are now in the same language, the problem of identifying comparable webpages is treated as near-duplicate detection. An index is built mapping n-grams to documents, and this index is used to find a bag-of-n-grams score for potentially comparable documents. The computation is kept feasible by only creating index entries for rare n-grams.

Ture and Lin (2012) used cross-lingual information retrieval techniques to find comparable document pairs in Wikipedia. While Wikipedia already provides annotated comparable document pairs through interwiki links, the authors consider all possible German-English article pairs as potentially containing comparable data.

2.2.2.2 Finding Parallel Sentences

Once comparable document pairs have been identified, most comparable corpora extraction methods will independently judge each sentence pair as parallel or non-parallel. Since there is often a very large amount of document pairs and thus potential sentence pairs, filters are used to prune out sentence pairs that are highly unlikely to be parallel. For example, Munteanu and Marcu (2005) used a sentence length filter to remove sentence pairs where one sentence was more than twice as long as the other. In addition, they used a word overlap filter based on the bilingual dictionary used to

find candidate document pairs.

Given a filtered set of sentence pairs, more expensive methods of scoring sentence pairs can be used. Munteanu and Marcu (2005) use a MaxEnt binary MaxEnt classifier to ultimately determine whether or not a sentence pair is parallel. The classifier is trained on parallel data and makes used of features which are mostly based on word alignments. Others Fung and Cheung (2004a,b); Tillmann (2009); Tillmann and Xu (2009) use a single score for sentence pairs based on either a word alignment model or bag-of-words similarity after projection through a bilingual lexicon, and tune a threshold on held out data.

Bibliography

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.

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.

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 trans-

lation. 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.

BIBLIOGRAPHY

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.

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.

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.

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

BIBLIOGRAPHY

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.

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.

BIBLIOGRAPHY

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.