Creating bilingual lexicon

King Fahd University of Petroleum and Minerals

College of Computer Science and Engineering

ICS 482

Presentation Report

Student Name: Hassan S. Al-Ayesh
ID: 224642

Summary:

Creating bilingual lexicon needs some processing on pairs of parallel documents of different languages. First, Some Processing sequence needs to be implemented for getting some parallel documents from the Internet. Next, "preprocessing" phase will be occurring. After that, either one of two defined algorithms or both of them must be processed. First Algorithm needs little time for processing but cannot handle group of words. The second algorithm is exactly the reverse of the first. The effect of stemming on the algorithms is a little increase in lexicon creation algorithms.

References:

1- Aljalyl and Frieder, 2002 M. Aljalyl and O. Frieder, On Arabic search: Improving the retrieval effectiveness via a light stemmer approach. In: K. Kalpakis *et al.*, Editors, *CIKM 2002: Proceedings of the Eleventh International Conference on Information and Knowledge Management (McLean, VA, Nov. 2002)*, ACM, New York (2002), pp. 340–347.
2- Brown et al., 1993 P.F. Brown, S.D. Pietra, V.D. Pietra and D.R. Ercer, The mathematics of statistical machine translation: parameter estimation, *Computational Linguistics* **19** (1993) (2).
3- Chen and Gey, 2001 Chen, A., & Gey, F. (2001). Translation term weighting and combining translation resources in cross-language retrieval. In *The Tenth Text REtrieval Conference (TREC 2001) Gaithersburg, Maryland* (pp. 529–533).
4- Chen and Gey, 2002 Chen, A., & Gey, F. (2002). Building an arabic stemmer for information retrieval. In *The Eleventh Text Retrieval Conference Gaithersburg, Maryland (TREC 2002)* (pp. 19–22).
5- Cherry and Lin, 2003 Cherry, C., & Lin, D. (2003). A probability model to improve word alignment. In *Proceedings of ACL-2003, Sapporo, Japan 2003* (pp. 88–95).
6- Darwish and Oard, 2002 Darwish, K., & Oard, D. W. (2002). Evidence combination for arabic–english retrieval. In *The Eleventh Text Retrieval Conference (TREC 2002), Gaithersburg, Maryland 2002*.
7- Déjean et al., 2002 Déjean, H., Gaussier, É., & Sadat, F. (2002). Bilingual terminology extraction: an approach based on a multilingual thesaurus applicable to comparable corpora. In *Proceedings of the 19th International Conference on Computational Linguistics COLING 2002* (pp. 218–224).
8- Djoerd, 1998 Djoerd, H. (1998). Multilingual domain modeling in twenty-one: automatic creation of a bi-directional translation lexicon from a parallel corpus. In P.-A. Coppen, H. van Halteren, & L. Teunissen (Eds.), *Proceedings of the eighth CLIN meeting 1998* (pp. 41–58).
9- Elamed, 1996 Elamed, D. I. (1996). Automatic construction of clean broad-coverage translation lexicons. In *Proceedings of the Second Conference of the Association for Machine Translation in the Americas (AMTA-96). Montreal 1996*.
10- Fattah et al., 2004 M.A. Fattah, F. Ren and K. Shingo, Internet archive as a source of bilingual dictionary, *Proceedings of International Conference on Information Technology: Coding and Computing (ITCC 2004)*, IEEE Computer Society, Las Vegas, Nevada (2004), pp. 298–302

11- Franz and McCarley, 2002 Franz, M., & McCarley, J. S. (2002). Arabic information retrieval at IBM. In *The Eleventh Text Retrieval Conference (TREC 2002), Gaithersburg, Maryland 2002*.
12- Hiemstra, 1997 Hiemstra, D. (1997). Deriving a bilingual lexicon for cross language information retrieval. In *Proceedings of Gronics 1997* (pp. 21–26).
13- Hiemstra et al., 1997 Hiemstra, D., Jong, F. M. G., & Kraaij, W. (1997). Domain specific lexicon acquisition tool for cross- language information retrieval. In *Proceedings of RIAO'97 Conference on Computer-Assisted Searching on the Internet 1997* (pp. 255–266).
14- Julapalli and Dhond, 2003 Julapalli, M., & Dhond, S. (2003). Word alignment in bilingual parallel corpora. CS224N/Ling237 Final Projects 2003, Spring 2002/2003.
15- Kay and Roscheisen, 1993 M. Kay and M. Roscheisen, Text-Translation alignment, *Computational Linguistic* **19** (1993) (1), pp. 121–142.
16- Lafourcade, 1997 Lafourcade, M. (1997). Multilingual dictionary construction and services case study with the Fe projects. In *PACLING'97—Meisei University—Ohme, Tokyo, Japan 1997* (pp. 171–181).
17- Larkey et al., 2003 Leah.S. Larkey, M.E. Connell and N.A. Jaleel, Hindi CLIR in thirty days, *ACM Transactions on Asian Language Information Process* **2** (2003) (2), pp. 130–142.
18- Lee et al., 2003 Lee, Y., Papineni, K., Roukos, S., Emam, O., & Hassan, H. (2003). Language model based arabic word segmentation. In *Proceedings of ACL-2003, Sapporo, Japan 2003* (pp. 399–406).
19- McEwanl et al., 2002 McEwanl, C. J. A., Ounis, I., & Ruthven, I. (2002). Building bilingual dictionaries from parallel web documents. LNCS, Spring 2002, 303–323.
20- Nie et al., 1999 Nie, J. Y., Simard, M., Isabelle, P., & Durard, R. (1999). Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the web. In *Proceedings of the 22nd International Conference on Research and Development in Information Retrieval (ACM SIGIR'99), Berkeley 1999* (pp. 74–81).
21- Niessen et al., 1998 Niessen, S., Vogel, S., Ney, H., & Tillmann, C. (1998). A DP-based search algorithm for statistical machine translation. In *COLING-ACL '98: Annual Conf. of the Association for Computational Linguistics and 17th Int. Conf. on Computational Linguistics, Montreal, Canada, August 1998* (pp. 960–967).
22- Och and Ney, 2003 F.J. Och and H. Ney, A systematic comparison of various statistical alignment models, *Computational Linguistics* **29** (2003) (1), pp. 19–51.
23- Och et al., 1999 Och, Josef, F., Tillmann, C., & Ney, H. (1999). Improved alignment models for statistical machine translation. In *Proc. of the Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, University of Maryland, College Park, MD, June 1999* (pp. 20–28).
24- Resnik and Smithy, 2003 Resnik, P., & Smithy, N. A. (2003). The web as a parallel corpus. University of Maryland technical report UMIACS-TR-2002-61 (also listed by technical report numbers CS-TR-4381 and LAMP-TR-089), July 2002. (Revised version to appear in Computational Linguistics 29(3), September 2003).
25- Rogati et al., 2003 Rogati, M., McCarley, S., & Yang, Y. (2003). Unsupervised learning of arabic stemming using a parallel corpus. In *Proceedings of ACL-2003, Sapporo, Japan 2003* (pp. 391–398).
26- Sadat et al., 2003 Sadat, F.,Yoshikawa, M., & Uemura, S. (2003). Learning bilingual translations from comparable corpora to cross-language information retrieval: hybrid statistics-based and linguistics-based approach. In *Proceedings of the Sixth International Workshop on Information Retrieval with Asian Languages IRAL, Sapporo, Japan 2003*.
27- Tanaka and Umemura, 1994 Tanaka, K., & Umemura, K. (1994). Construction of a bilingual dictionary intermediated by a third language. In *Proceedings of the 15th International Conference on Computational Linguistics*.
28- Tillmann et al., 1997 Tillmann, C.,Vogel, S., Ney, H., & Ubiaga, A. (1997). A DP-based search using monotone alignments in statistical translation. In *Proc. 35th Annual Conf. of the Association for Computational Linguistics, Madrid, Spain, July 1997* (pp. 289–296).

29- Utiyama and Isahara, 2003 Utiyama, M., & Isahara, H. (2003). Reliable measures for aligning Japanese–English news articles and sentences. In *Proceedings of ACL-2003, Sapporo, Japan 2003* (pp. 72–79).
30- Venugopal et al., 2003 Venugopal, A., Vogel, S., & Waibel, A. (2003). Effective phrase translation extraction from alignment models. In *Proceedings of ACL-2003, Sapporo, Japan 2003* (pp. 319–326).
31- Vogel et al., 1996 Vogel, S., Ney, H., & Tillmann, C. (1996). HMM-based word alignment in statistical translation. In *COL1NG '96: The 16th Int. Conf. on Computational Linguistics, Copenhagen, August 1996* (pp. 836–841).
32- Xu et al., 2001 Xu, J., Fraser, A., & Weischede, R. (2001). TREC 2001 Cross-lingual retrieval at BBN. In *The Tenth Text REtrieval Conference (TREC 2001), Gaithersburg, Maryland 2001* (pp. 68–77).
33- Xu and Weischedel, 2000 Xu, J., & Weischedel, R. (2000). TREC-9 cross-lingual retrieval at BBN. In *The Ninth Text REtrieval Conference (TREC 9), Gaithersburg, Maryland 2000* (pp. 106–116).

Obstacles & Problems:

The biggest problem that has been faced is fining a good resource for the presentation.

Skills Learned:

1- Presenting papers with confidence & for long time.
2- Preparing a good presentation with good speech.

Recommendations:

- Prepare your self for presentation and read it two or more times.

True/False Questions:

1- First Algorithm is less efficient than second Algorithm. ( )
2- Both First & Second algorithms could be used in the system. ( )
3- Stemmer will decrease the accuracy of the system. ( )