Local Context Selection for Aligning Sentences in Parallel Corpora

Ergun Biçici

ebicici@ku.edu.tr

Koç University, Istanbul, Turkey

www.ku.edu.tr

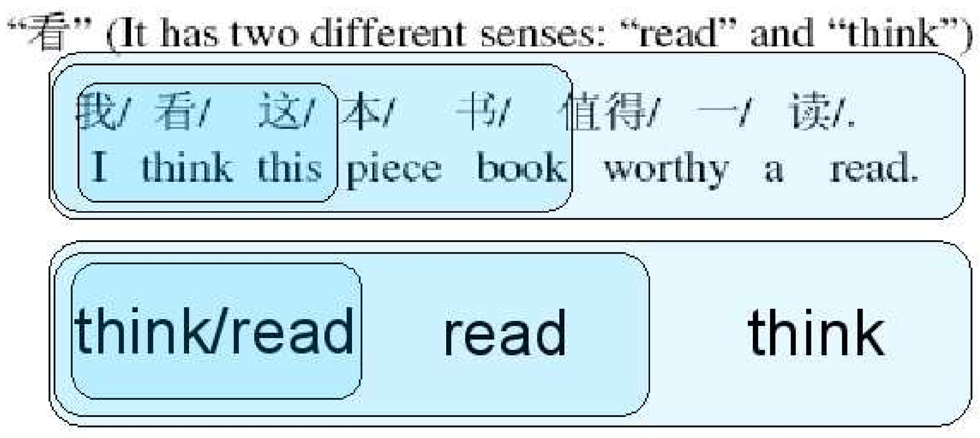
August 22, 2007

Abstract

A novel language-independent context-based sentence alignment technique, which uses the context of sentences and Zipfian word vectors, is presented. Alternatives for local context models examined and a demonstration of better performance when compared with prominent sentence alignment techniques is given. The local context for a pair of set of sentences which maximizes the correlation is dynamically selected. Our system performs 1.1951 to 1.5404 times better in reducing the error rate in alignment accuracy and coverage.

Motivation

- Sentence alignment: mapping the sentences of two given parallel corpora which are known to be translations of each other.
- Mappings are not necessarily 1-to-1, monotonic, or continuous.
- Dynamic nature of the context is noticed for many NLP tasks, including WSD [1]:



Zipfian Word Vectors

- Zipf's Law: "a few words occur frequently while many occur rarely" [2].
- Zipfian Word Vector (ZWV) turns a given sentence to bins where each bin contains the number of words with similar frequencies in the given corpus.
- Thus, for a single sentence as in:

  $S =$ `" big brother is watching you " , the caption beneath it ran .`,

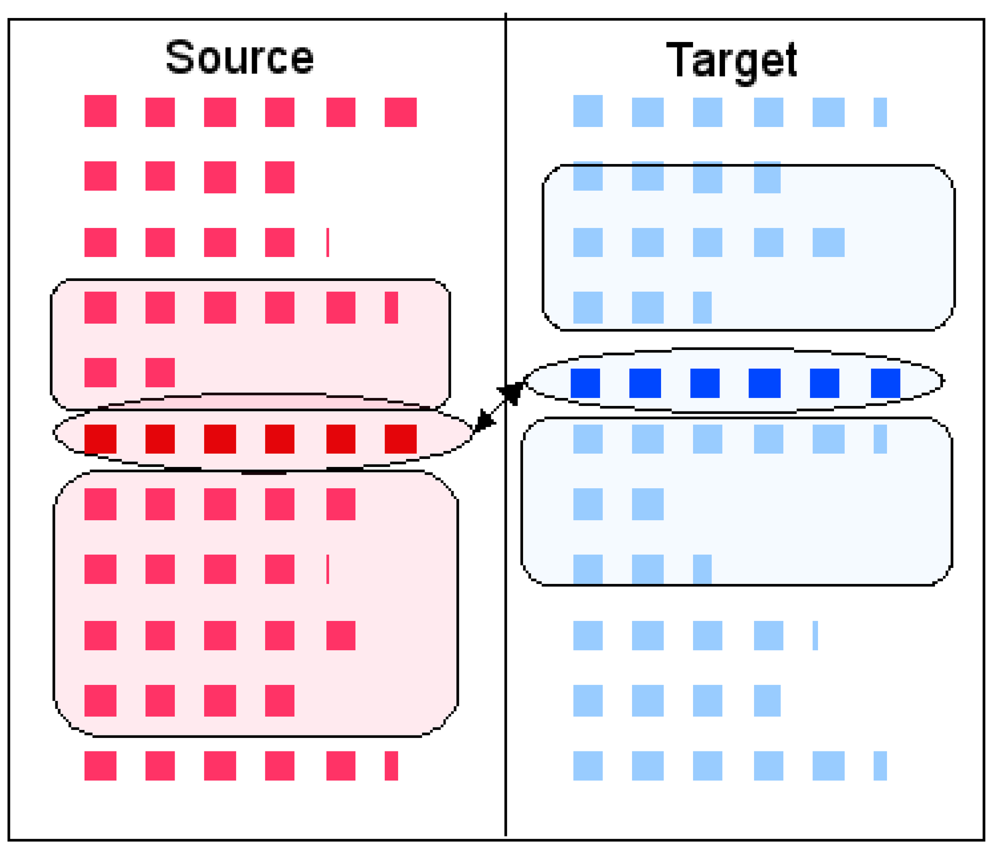
  the Zipfian word vector becomes:

  $$\mathrm{ZWV(S)} = [14, 1, 3, 0, 1, 3, 2, 0, 1, 1, 2],$$

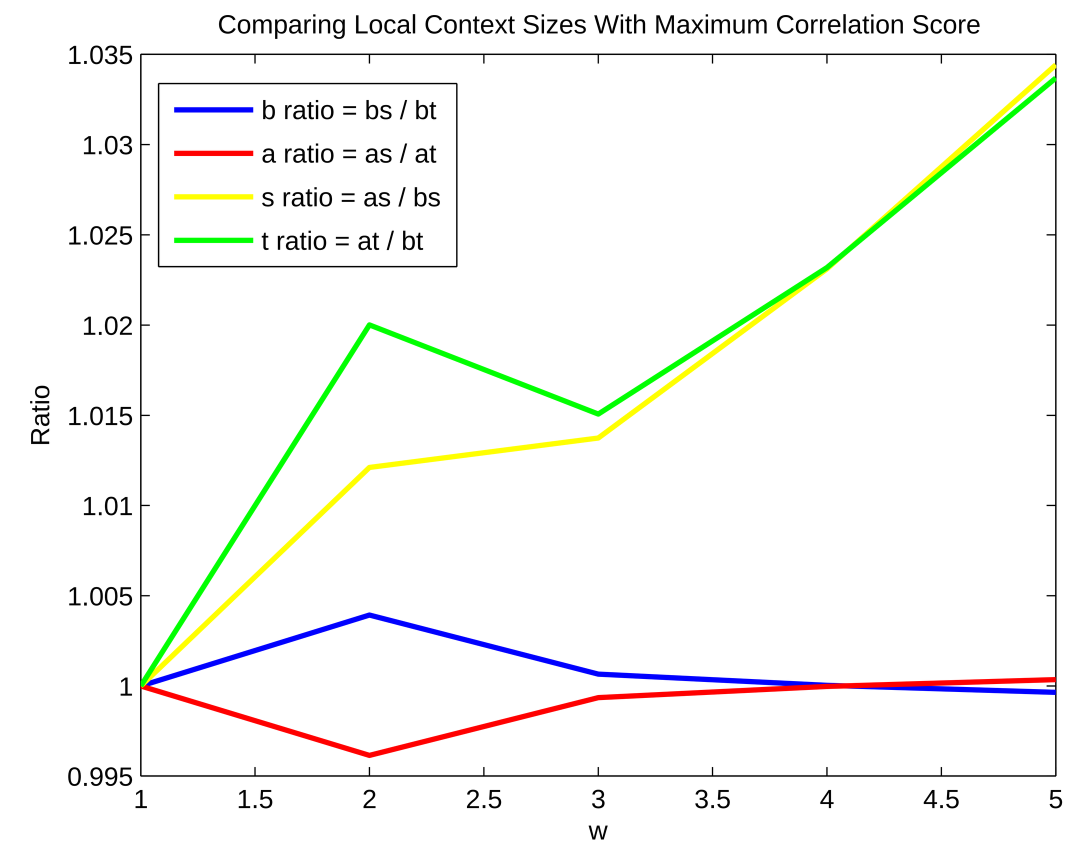
  where the sentence length in the number of tokens is added to the beginning as well.
- For each set of sentences we create the *Zipfian word matrix*, which is the concatenation of ZWVs surrounding S based on S's local context, which contains at most $2 \times w + 1$ rows for a given window size of w.
- Weight decaying 2D Pearson correlation coefficient is used for comparing candidate T's.

Context in Sentence Alignment

Example sentence alignment scenario is given below:



- The local contexts for the scenario are $(2, 4)$ and $(3, 3)$ respectively.
- *Full local context search* considers all w^3 possibilities for (b_s, a_s) and $(b_t, b_s + a_s - b_t)$.
- *Symmetric local context search* considers w^2 symmetric possibilities: $(b_s = b_t, a_s = a_t)$
- From $\mathcal{C}$ local context configurations, choose **maximum**, **average**, or **average top** k?



- The sentences that come before have a larger role in determining the context?
- In nearly two thirds of $\mathcal{C}$, the local context sizes for S and T are exactly the same.

Dataset	b	a	Increase
Bulgarian	1.9564	1.9936	1.9%
Czech	1.9610	1.9961	1.8%
Estonian	1.9563	2.0129	2.9%
Hungarian	1.9723	1.9964	1.2%
Lithuanian	1.9720	2.0246	2.7%
Latvian	1.9667	2.0043	1.9%
Romanian	1.9275	1.9688	2.1%
Serbo-Croatian	1.9424	1.9755	1.7%
Slovene	1.9486	1.9765	1.4%

TABLE 1: Local symmetric context sizes per language - English pairs

Results

	Sentence Alignment Accuracy					
Language	**hunalign**	**Moore**	**static**	**maximum**	**average**	**average top 5**
Bulgarian	**96.74 / 3.26**	96.09 / 3.91	96.09 / 3.91	95.77 / 4.23	**96.74 / 3.26**	**96.74 / 3.26**
Czech	96.14 / 3.86	95.82 / 4.18	**96.78 / 3.22**	**96.78 / 3.22**	**96.78 / 3.22**	**96.78 / 3.22**
Estonian	**99.68 / 0.32**	98.39 / 1.61	98.39 / 1.61	99.04 / 0.96	98.39 / 1.61	99.04 / 0.96
Hungarian	87.86 / 12.14	88.96 / 11.04	92.98 / 7.02	91.30 / 8.70	**93.98 / 6.02**	91.64 / 8.36
Latvian	95.71 / 4.29	92.74 / 7.26	96.70 / 3.30	96.70 / 3.30	96.70 / 3.30	**97.03 / 2.97**
Lithuanian	88.44 / 11.56	82.31 / 17.69	92.18 / 7.82	**92.52 / 7.48**	91.84 / 8.16	**92.52 / 7.48**
Romanian	89.86 / 10.14	**95.27 / 4.73**	91.22 / 8.78	90.54 / 9.46	91.22 / 8.78	92.23 / 7.77
Serbo-Croatian	**98.70 / 1.30**	97.08 / 2.92	97.73 / 2.27	98.05 / 1.95	97.73 / 2.27	97.73 / 2.27
Slovene	97.70 / 2.30	97.04 / 2.96	98.68 / 1.32	98.36 / 1.64	**99.34 / 0.64**	98.36 / 1.64

TABLE 2: Sentence alignment accuracy per English - language alignments

	Sentence Alignment Coverage					
Language	**hunalign**	**Moore**	**static**	**maximum**	**average**	**average top 5**
Bulgarian	95.34 / 4.66	94.86 / 5.14	95.18 / 4.82	94.86 / 5.14	**95.99 / 4.01**	**95.99 / 4.01**
Czech	94.92 / 5.08	95.24 / 4.76	**96.35 / 3.65**	**96.35 / 3.65**	**96.35 / 3.65**	**96.35 / 3.65**
Estonian	**99.52 / 0.48**	98.08 / 1.92	98.08 / 1.92	98.88 / 1.12	98.08 / 1.92	98.88 / 1.12
Hungarian	84.30 / 15.70	85.90 / 14.10	91.51 / 8.49	89.10 / 10.90	**92.63 / 7.37**	89.42 / 10.58
Latvian	92.65 / 7.35	90.26 / 9.74	95.37 / 4.63	95.37 / 4.63	95.37 / 4.63	**95.69 / 4.31**
Lithuanian	84.85 / 15.15	79.15 / 20.85	90.23 / 9.77	**90.72 / 9.28**	89.90 / 10.10	**90.72 / 9.28**
Romanian	86.79 / 13.21	**93.64 / 6.36**	89.72 / 10.28	89.07 / 10.93	89.72 / 10.28	90.86 / 9.14
Serbo-Croatian	**97.75 / 2.25**	96.46 / 3.54	97.27 / 2.73	97.59 / 2.41	97.27 / 2.73	97.27 / 2.73
Slovene	95.81 / 4.19	95.64 / 4.36	98.06 / 1.94	97.58 / 2.42	**98.71 / 1.29**	97.58 / 2.42

TABLE 3: Sentence alignment coverage per English - language alignments

Conclusions

- Provided formalizations of context for the sentence alignment task.
- Introduced Zipfian word vectors, which effectively presents an order-free representation of the distributional properties of a given sentence.
- Defined 2D weight decaying correlation for calculating the similarities between sentences.
- Our system dynamically selects the local context for a pair of set of sentences which maximizes the correlation.
- The system performs 1.1951 to 1.5404 times better in reducing the error rate in alignment accuracy and coverage.

References

[1] Xiaojie Wang. Robust utilization of context in word sense disambiguation. In Anind Dey, Boicho Kokinov, David Leake, and Roy Turner, editors, *Modeling and Using Context: 5th International and Interdisciplinary Conference*, pages 529541. Springer-Verlag, Berlin, 2005.

[2] George Kingsley Zipf. The meaning-frequency relationship of words. *The Journal of General Psychology*, 33:251256, 1945.

Acknowledgments: The research reported here was supported in part by the Scientific and Technological Research Council of Turkey (TUBITAK). The author would like to thank Deniz Yuret for helpful discussions and for guidance and support during the term of this research as well as Nokia for their generous support for our presence at Context 2007.