Application of Clause Alignment for Statistical Machine Translation

Svetla Koeva, Borislav Rizov, Ivelina Stoyanova, Svetlozara Leseva, Angel Genov, Rositsa Dekova, Ekaterina Tarpomanova, Tsvetana Dimitrova and Hristina Kukova

Department of Computational Linguistics, IBL, Bulgarian Academy of Sciences

Introduction

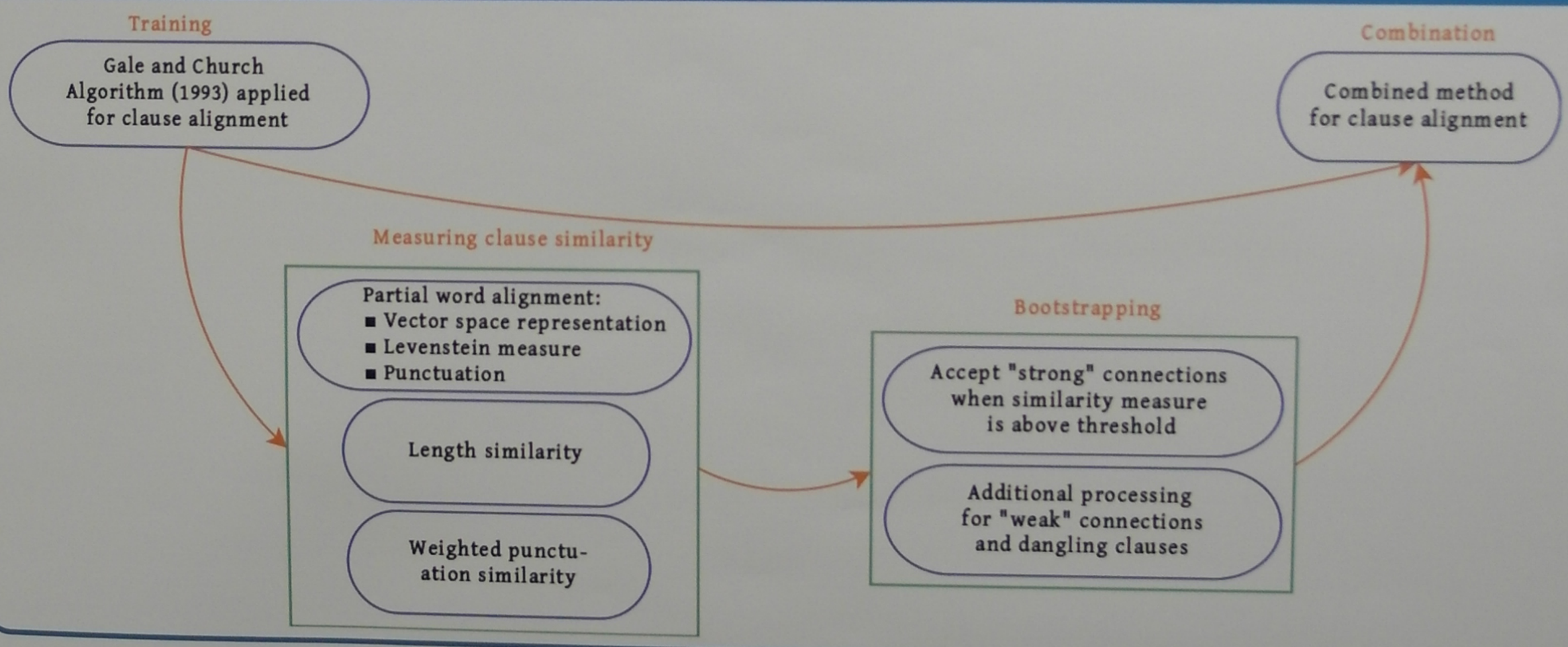
The paper presents a new resource light flexible method for clause alignment combining Gale-Church algorithm with internally collected textual information.

Clause alignment can provide improved training data for the purposes of Statistical Machine Translation (SMT). A series of experiments with Moses demonstrate ways to modify the parallel training resources and the effects on translation quality by clause alignment and clause reordering.

Motivation of the study

- An efficient way to overcome the problem of sentence length and complexity is to process the clauses in a similar way as sentences.
- Differences in word order and phrase structure across languages can be better captured at clause rather than at sentence level.
- Monolingual and parallel text processing whose scope are the clauses may significantly improve syntactic parsing, automatic translation, etc.
- Phrase-based and the syntax-based SMT systems do not perform well on long and (syntactically) complex sentences.
- Experiments demonstrate the effect of syntactic information (reordering the clauses within the source language sentences) on the performance of the SMT system.

Combined method for clause alignment



Experiments with Moses

The training corpus:

- 27,408 aligned sentence pairs;
- BG: 382,950 tokens; EN: 409,757 tokens;
- Semi-automatically split into clauses and automatically aligned;
- Domains – Administrative, Fiction, News, Science, Subtitles.

The series of experiments conducted with Moses showed possible applications of the clause alignment method for training an SMT system, enhanced with linguistic information.

The experiments:

Baseline – trained on aligned sentence pairs.

Experiment 1 – trained on aligned clause pairs.

Experiment 2 – trained on reordered sentence pairs. Reordering is applied within the sentence and clauses in the SL are reordered to match the order of corresponding clauses in the TL. Affects about 7% of sentences.

Results and conclusion

	BLEU	Diff.
Baseline	16.99	N/A
Experiment 1	17.10	+0.11
Experiment 2	17.12	+0.13

- Small consistent improvement in the BLEU score by training performed on parallel data enhanced with syntactic information – aligned clause pairs or sentences with reordered clauses.
- Results are inconclusive both with respect to whether the improvement is stable and which of the two methods – using clause aligned pairs or reordered sentences – performs better.