**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**

- **Training**
  - Gale and Church Algorithm (1993) applied for clause alignment

- **Measuring clause similarity**
  - Partial word alignment:
    - Vector space representation
    - Levenshtein measure
    - Punctuation
  - Length similarity
  - Weighted punctuation similarity

- **Combination**
  - Combined method for clause alignment
  - Bootstrapping
    - Accept "strong" connections when similarity measure is above threshold
    - Additional processing for "weak" connections and dangling clauses

**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;

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**

<table>
<thead>
<tr>
<th></th>
<th>BLEU</th>
<th>Diff.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>16.99</td>
<td>N/A</td>
</tr>
<tr>
<td>Experiment 1</td>
<td>17.10</td>
<td>+0.11</td>
</tr>
<tr>
<td>Experiment 2</td>
<td>17.12</td>
<td>+0.13</td>
</tr>
</tbody>
</table>

- 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.