Machine Translation

Clause Restructuring for Statistical Machine Translation

Summary

Aqeel Labash

1 Introduction

The idea of the paper is to simply parse the source language and create a tree out of it then apply some transformation over the words. The transformation will make the source sentence closer to the destination sentence by mean of words location.

2 Background

They spoke about IBM model implicitly (which derive the word to word alignment). Then about Koehn which is the phrase system they used in the paper. Koehn depends on the output of IBM model. Then generate phrase to phrase pairs. They also spoke about baseline system that generates N-best translation. The syntactic features then pick the winner.They also mention information about methods used to reorder specific cases (`NOUN de NOUN`). Simple German language structure (finite , infinitive verbs) and the position of each state and how much german language is free context was covered there as well.

3 Algorithm

To explain the algorithm I use the examples that they used in the paper and work with it step by step.

3.1 Initialization

This is not part of there reordering methodology but it's essential for their algorithm to work. Basicly what they do here is to build a tree of **annotated** words. For example the sentence `wir fordern das Paraesidium auf` will become :

```
  S PREP-SB       Wir
   VVFIN-HD     fordern
  NP-OA ART       das
         NN   Paraesidium
  PTKVZ-SVP       auf
```

after the text being annoteated in the previous forum the algorithm start.

3.2 Verb initial

In verb phrase 1, find head phrase and move it into initial position 2 for example sentence `Ich werde Ihnen die entsprechenden anmerkungen aushaendigen damit Sie das eventuell bei der Abstimmung uerbernehmen koennen`
Become: `Ich werde aushaendigen Ihnen die entsprechenden anmerkungen damit Sie uerbernehmen das eventuell bei der Abstimmung koennen`

3.3 Verb 2nd

Any sbuordinate claus with `S-..→KOUS,PREL,PWS,PWAV` label the head of that claus would be moved to follow the complementizer. For example :
`damit Sie uerbernehmen das eventuell bei der Abstimmung koennen`
Become: `damit koennen Sie uerbernehmen das eventuell bei der Abstimmung`

3.4 Move Subject

For `S` labeled clausses the subject3 would be moved to directly precede the head4. So the sentence we processed in (3.2) will become `damit Sie koennen uerbernehmen das eventuell bei der Abstimmung`

3.5 Particles

Move the the particle to directly precede the verb. if we used 3.1 example it will become `Wir auf fordern das Praesidium`

3.6 Infinitives

This stage has two steps: **1**- remove all `VP` labeled nodes. **2-** if the clause contained finite verb, infinitve verb and an argument (subject or object) and the argumet where between the finite and infinitve verb the infinitve verb then the infinitive verb would follow the finite verb. Example: `Wir konnten es nicht einreichen mehr rechtzeitig`
Become: `Wir konnten einreichen es nicht mehr rechtzeitig`

3.7 Negation

Move negative particlae to follow the finite verb when finite and infinitive verbs exist in same sentence. Example : `Wir konnten nicht einreichen es mehr rechtzeitig`.

4 Experments and results

The experment used the Europarl corpus. `751,088` sentence pairs with `15,256,792` German word and `16,052,269` English word. Test done on 2000 sentence with average 28 word per sentence.The accuracy measurement used were BLEU. The BLEU score was`25.2%` compared to `26.8%` using there reordering way.

4.1 Human Translation Judgements

Two annotaters viewed 100 randomly picked translations with length between 10-20 words per sentence. The annotater have to select what they prefered as shown in Table1

	Annotator2		
Annotator1	R	B	E
R	33	2	5
B	2	13	5
E	9	4	27

Table 1: Annotator 1,2 decisions, R: reordered,B: baseline, E:both same

4.2 Statistical Significance

Under the baseline system 52.85% of test cases were improved. 36.4% were worse. 10.75% had same quality as before.**Note:**I think they meant under the **reordered system** here depending on there definition but got to stick with there text. Also there 95% confidence that the reordered system will perform $[56.9\%, 61.5\%]$ better than the baseline.

My Notes:

Annotation Problem: one of the things that I noticed while reading this paper that they didn't mention anything about annotating the text. While as far as I know it's not an easy problem and it's an ongoing research to use machine learning for it (maybe the book I read was an old book with outdated info about this case). **Statistics Study**

Referenced info

1. Verb phrase could be identified with special annotation.
2. didn't know what they mean by initial position
3. left most child with label `-SB , PPEREP`
4. left most child with label `-HD`
5. c_+ count of ordered is better than baseline , c_- count of baseline better , c_0 count of both same quality.