Automatic Translation Error Analysis

Mark Fishel
Dept. of Computer Science
University of Tartu, Estonia
fishel@ut.ee

Abstract

1 Introduction

[Most efforts on MT eval concentrated on producing a single score (BLEU, NIST, METEOR, TER, SemPOS, LRscore, ad inf). While that's convenient for comparison, it is not informative.]

[manual evaluation does scoring (HTER, fluency/adequacy, rank) and some analysis (Vilar et al. 2006). **TODO** other examples.]

[we introduce a method of automatic analysis of translation errors. It only requires the source, reference and hypothesis translations. The whole thing is language independent, but is capable of taking additional information into account, such as linguistic analysis (lemmatizing, PoS tagging, synonym detection), training sets or dictionaries, etc.]

[in this work we tune our method to mimick the error taxonomy of Vilar et al.]

[Evaluated: a) in comparison with manual analysis of 4 En-Cz translations and b) by comparing the weaknesses of some state-of-the-art statistical systems (Moses, cdec, Google) and comparing the result with what we expected (like better order modelling in one of the systems, etc.); any languages]

2 Related Work

3 Method Description

[Word alignment between hypothesis and reference, error detection and classification, error summarization.]

3.1 Alignment

[The first requirement and the first step – align the words. Whenever 1 word is aligned to several, the several are treated as a single group (like in Tiedemann paper in COLING 2004). This automatically adds the only restriction on alignment – no aligning of one word to two or more **non-adjacent** words.]

[Can be any kind of alignment, by METEOR, GIZA++, or anything (**TODO** refs).]

[Aligning same-language sentences seems to be a trivial task; the only thing that makes it difficult is ambiguity: many repeating tokens ("the", punctuation, ...), synonyms, same-lemma-wrong-form examples, etc.]

[Here – HMM-model-like alignment, hyp words are the observed variables, ref words – hidden variables. Instead of learning the emission/transition probabilities, these are hand-crafted. Therefore the resulting model should have the advantages of HMM alignment without the necessity to learn the models (good for applying to a small set of translations).]

[Emission probability depends on the number of the same words in the hypothesis; for a hyp word $w_i^{(h)}$ that occurs only once

$$p_{emit}(w_i^{(h)}|w_{a_i}^{(r)}) = [w_i^{(h)} = w_{a_i}^{(r)}];$$

for other words (occurring many times)

$$p_{emit}(w_i^{(h)}|\emptyset) = \varepsilon,$$

$$p_{emit}(w_i^{(h)}|w_{a_i}^{(r)}) = \frac{(1-\varepsilon)\cdot[w_i^{(h)} = w_{a_i}^{(r)}]}{|\{w : w \in \text{hyp}, w = w_i^{(h)}\}|}.$$

This allows repeating words to remain unaligned to make way for other, potentially better alignments of the same word, while always aligning unique words to their counterpart.]

[The transition probabilities stimulate aligning the current pair in parallel to the previously produced pair by penalizing the distance between the previous and the current reference word:

$$p_{trans}(w_{a_i}^{(r)}|w_{a_{i\_}}^{(r)}) \sim \exp(-b \cdot |a_i - a_{i\_} - 1|),$$

where $a_{i\_}$ is the index of the latest non-NULL alignment in the alignment $\mathbf{a}$.]

[We do alignment based on lemmas, in order to detect same-lemma-wrong-form translations; alternative is detecting synonyms (synosets?) and using them for alignment to support detecting synonym translations.]

[**TODO** – describe here or in rel. work sct, how METEOR and TER do alignment, and whether that's better or worse than our method.]

3.2 Detecting Lexical Errors

[Using the alignment we detect and classify translation errors (correspondence errors):]

- unaligned words in the reference = missing words
  - can be further classified into punctuation and content/filler words, using PoS-tagging
- unaligned words in the hypothesis = extra words
  - also further classified (punctuation, content/filler word)
  - if present in the source sentence – then it's an untranslated word
- aligned but different word form = same-lemma-wrong-form translation

3.3 Detecting Order Errors

[The aligned words and word groups are in a 1-to-1 correspondence; this can be used to calculate Hamming distance, Kendall's τ distance, Ulam's distance (Birch,Blunsom,Osborne@MT Journal, 2010), Spearman's rank correlation coefficient, etc.]

[Here we want to detect misplaced words and word groups. For that we do a breadth-first search for fixing the order in the aligned hypothesis words. The weighed DAG is created like this:]

- one node per every permutation
- there's an arc iff the target node permutation differs from the source permutation by two adjacent symbols, whereas the relative order of the two symbols is wrong in the source and correct in the target
- the arc weight is generally equal to 1; in order to enable block shifts, the weight is 0 only if the current arc continues shifting a symbol to the right or to the left

[**TODO**: currently missing, but have to group adjacent word alignments into word groups – this will make the unscrambling faster, and the groups can be treated as phrazes for order error detection; Ondrej, Dan – does the latter make sense?]

3.4 Summarization

[Can be done on many levels:]

- no summarization – inspecting translation errors in every sentence = Addicter! :)
- summarization la (Vilar et al. 2006) – ratio of missing/untranslated/extra/wrong-form/misplaced words
- linear combination of the error type ratios – score!

4 Experiments and Results

TODO:

- apply Morče to the 4 En-Cz translations from WMT09
- apply the introduced analysis to it
- compare results to Ondrej's MT Journal article analysis

5 Conclusions

```
Total sentences:       2656
Total ref words:      23536
Total hyp words:      20432

Missing ref words:
       number:      30 ( 0.13% of ref)
        punct:     183 ( 0.78% of ref)
      content:    4439 (18.86% of ref)
       filler:    5254 (22.32% of ref)

        total:    9906 (42.09% of ref)

Incorrect hyp words:
   wrong form:     621 ( 3.04% of hyp)
 untranslated:     507 ( 2.48% of hyp)

extra, number:      27 ( 0.13% of hyp)
 extra, punct:     161 ( 0.79% of hyp)
extra, content:    3721 (18.22% of hyp)
extra, filler:    2384 (11.67% of hyp)

        total:    7423 (36.33% of hyp)

Order similarity metrics
Spearman's rho: 0.930

Word order errors, by shift distance:
            1:     152 ( 0.74% of hyp)
            2:     207 ( 1.01% of hyp)
            3:     191 ( 0.93% of hyp)
            4:     156 ( 0.76% of hyp)
            5:     117 ( 0.57% of hyp)
            6:     103 ( 0.50% of hyp)
            7:      60 ( 0.29% of hyp)
           8+:     294 ( 1.44% of hyp)

        total:    1280 ( 6.26% of hyp)
```

Figure 1: Example of analysis, done for Google's Ee-En translation