Statistical NLP: Lecture 15

Statistical Alignment and Machine Translation

(Ch 13)

Overview

- MT is a difficult problem: translation programs available today do not perform very well.
- Different approaches to MT:
  - *Word for Word*
  - *Syntactic Transfer Approaches*
  - *Semantic Transfer Approaches*
  - *Interlingua*
- Most MT systems are a mix of probabilistic and non-probabilistic components, though there are a few completely statistical translation systems.

Overview (Cont'd)

- A large part of implementing an MT system [e.g., probabilistic parsing, word sense disambiguation] is not specific to MT.
- Nonetheless, parts of MT that are specific to it are: ***text alignment*** and ***word alignment***.
- ***Definition:*** In the sentence alignment problem, one seeks to say that some group of sentences in one language corresponds in content to some other group of sentences in another language. Such a grouping is referred to as a ***bead of sentences***.

Overview of the Lecture

- Text Alignment
- Word Alignment
- Fully Statistical Attempt at MT

Text Alignment: Aligning Sentences and Paragraphs

- ***Text alignment*** is useful for bilingual lexicography, MT, but also as a first step to using bilingual corpora for other tasks.
- Text alignment is not trivial because translators do not always translate one sentence in the input into one sentence in the output, although they do so in 90% of the cases.
- Another problem is that of ***crossing dependencies***, where the order of sentences are changed in the translation.

Different Approached to Text Alignment

- ***Length-Based Approaches***: short sentences will be translated as short sentences and long sentences as long sentences.
- ***Offset Alignment by Signal Processing Techniques***: these approaches do not attempt to align beads of sentences but rather just to align position offsets in the two parallel texts.
- ***Lexical Methods:*** Use lexical information to align beads of sentences.

Length-Based Methods I: General Approach

- ***Goal:*** Find alignment A with highest probability given the two parallel texts S and T: $argmax_A P(A|S, T)=argmax_A P(A, S, T)$
- To estimate the above probabilities, the aligned text is decomposed in a sequence of aligned beads where each bead is assumed to be independent of the others. Then $P(A, S, T) \approx \Pi_{k=1..K} P(B_k)$.
- The question, then, is how to estimate the probability of a certain type of alignment bead given the sentences in that bead.

Length-Based Methods II: Gale and Church, 1993

- The algorithm uses sentence length (measured in characters) to evaluate how likely an alignment of some number of sentences in L1 is with some number of sentences in L2.
- The algorithm uses a Dynamic Programming technique that allows the system to efficiently consider all possible alignments and find the minimum cost alignment.
- The method performs well (at least on related languages). It gets a 4% error rate. It works best on 1:1 alignments [only 2% error rate]. It has a high error rate on more difficult alignments.

Length-Based Methods II: Other Approaches

- ***Brown et al., 1991:*** Same approach as Gale and Church, except that sentence lengths are compared in terms of words rather than characters. Other difference in goal: Brown et al. Didn't want to align entire articles but just a subset of the corpus suitable for further research.
- ***Wu, 1994***: Wu applies Gale and Church's method to a corpus of parallel English and Cantonese text. The results are not much worse than on related languages. To improve accuracy, Wu uses lexical cues.

Offset Alignment by Signal Processing Techniques I : Church, 1993

- Church argues that length-based methods work well on clean text but may break down in real-world situations (noisy OCR or unknown markup conventions)
- Church's method is to induce an alignment by using ***cognates*** (words that are similar across languages) at the level of character sequences.
- The method consists of building a dot-plot, i.e., the source and translated text are concatenated and then a square graph is made with this text on both axes. A dot is placed at (x,y) when there is a match [4-gram char].

Offset Alignment by Signal Processing Techniques II: Church, 1993 (Cont'd)

- Signal processing methods are then used to compress the resulting plot.
- The interesting part in a dot-plot is called the ***bitext maps***. These maps show the correspondence between the two languages.
- In the bitext maps, there are faint, roughly straight diagonals corresponding to cognates.
- A heuristic search along this diagonal provides an alignment in terms of offsets in the two texts.

Offset Alignment by Signal Processing Techniques III: Fung & McKeown, 1994

- Fung and McKeown's algorithm works:
  - Without having found sentence boundaries.
  - In only roughly parallel text (with certain sections missing in one language)
  - With unrelated language pairs.
- The technique is to infer a small bilingual dictionary that will give points of alignment.
- For each word, a signal is produced, as an arrival vector of integer numbers giving the number of words between each occurrence of the word at hand.

Lexical Methods of Sentence Alignment I: Kay & Roscheisen, 1993

- Assume the first and last sentences of the texts align. These are the initial ***anchors***.
- Then, until most sentences are aligned:
  1. Form an ***envelope*** of possible alignments.
  2. Choose pairs of words that tend to co-occur in these potential partial alignments.
  3. Find pairs of source and target sentences which contain many possible lexical correspondences. The most reliable of these pairs are used to induce a set of ***partial alignments*** which will be part of the final result.

Lexical Methods of Sentence Alignment II: Chen, 1993

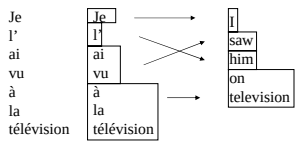
- Chen does sentence alignment by constructing a simple word-to-word translation model as he goes along.
- The best alignment is the one that maximizes the likelihood of generating the corpus given the translation model.
- This best alignment is found by using dynamic programming.

Lexical Methods of Sentence Alignment III: Haruno & Yamazaki, 1996

- Their method is a variant of Kay & Roscheisen (1993) with the following differences:
  - For structurally very different languages, function words impede alignment. They eliminate function words using a POS tagger.
  - If trying to align short texts, there are not enough repeated words for reliable alignment using Kay & Roscheisen (1993). So they use an online dictionary to find matching word pairs.

Word Alignment

- Align each word in one sentence with a word in the other sentence.
- Some words may align with null.



Word Alignment

- A common use of aligned texts is the derivation of bilingual dictionaries and terminology databases.
- This is usually done in two steps: First, the text alignment is extended to a word alignment. Then, some criterion, such as frequency is used to select aligned pairs for which there is enough evidence to include them in the bilingual dictionary.
- Using a χ^2 measure works well unless one word in L1 occurs with more than one word in L2. Then, it is useful to assume a one-to-one correspondence.
- Future work is likely to use existing bilingual dictionaries.

Fully Statistical MT I

- MT has been attempted using a ***noisy channel model***. Such a model requires:
- A Language Model
- A Translation Model (Translation Probabilities)
- A Decoder
- An evaluation of the model found that only 48% of French sentences were translated correctly. The errors were either incorrect decodings or ungrammatical decodings.

Fully Statistical MT II: Problems with the Model

- Fertility is Asymmetric
- Independence Assumptions
- Sensitivity to Training Data
- Efficiency
- No Notion of Phrases
- Non-Local Dependencies
- Morphology
- Sparse Data Problems.
- In summary, non-linguistic models are fairly successful for word alignments, but they fail for MT.

Evaluation of MT systems

- BLUE score
- NIST score
- Human judges