PARSER PROJECTION

Biplab Ch Das

Ravi Kumar

Source of Slides

Franco M. Luque

Ten (Or More) Minutes on Unsupervised Parsing

Supervised Parsing

Supervised Parsing is the problem of building a parser using a treebank.

Treebanks are corpuses of parsed sentences.

- A part of the treebank is used to train the parser.
- Another part is used to evaluate the parser.
- It can be seen as learning a function $f : X \rightarrow Y$ (the parser) by

knowing some points (x1 , f (x1)), . . . , (xn , f (xn)) (the treebank).

Unsupervised Parsing

What can we do by only knowing a set of points $\{x_1, \ldots, x_n\}$ of

the domain of f ?

The parser is trained with sentences.

- The evaluation is still done against a treebank.
- The set of syntactic categories is unknown.

The DMV+CCM Model

Developed by Dan Klein and Chris Manning in 2004.

Parses dependency trees that can be converted to binary bracketings.

Learn and parse from POS tags instead of words. Must be combined with a POS tagger to obtain a real parser.

Evaluated on English, German and Chinese treebanks, only with sentences of length ≤ 10.

Punctuation is not considered in order to emulate spoken language.

Combining the two Models and training

CCM Stands for Constituent Content Model

DMV stands for Dependency Model with Valence

Versions of the inside-outside algorithm for PCFG's [Lari and Young, 1990] can be used for running estimation maximation on both these probabilistic models.

But How can we use The parsed tree of other language

Alignment Problem!!

Lets consider deviate a bit from original problem.

Suppose we have a Parallel tree bank and we need to find the alignment between sentences given in two different languages.

Solution:

We can get the alignment using dependencies and constituents for Tree-Based Alignment.

(ref: Dependencies vs. Constituents for Tree-Based Alignment by Daniel Gildea)

But that was not our problem

Lets see word problem.

Word alignment was Discussed in class.

There are other approaches to get word alignment using the hybrid approach provided in [A hybrid approach to align sentences and words in English-Hindi parallel corpora, Niraj Aswani and Robert Gaizauskas]

Lets see an abstract view of the algorithm.

The first part Dictionary lookup

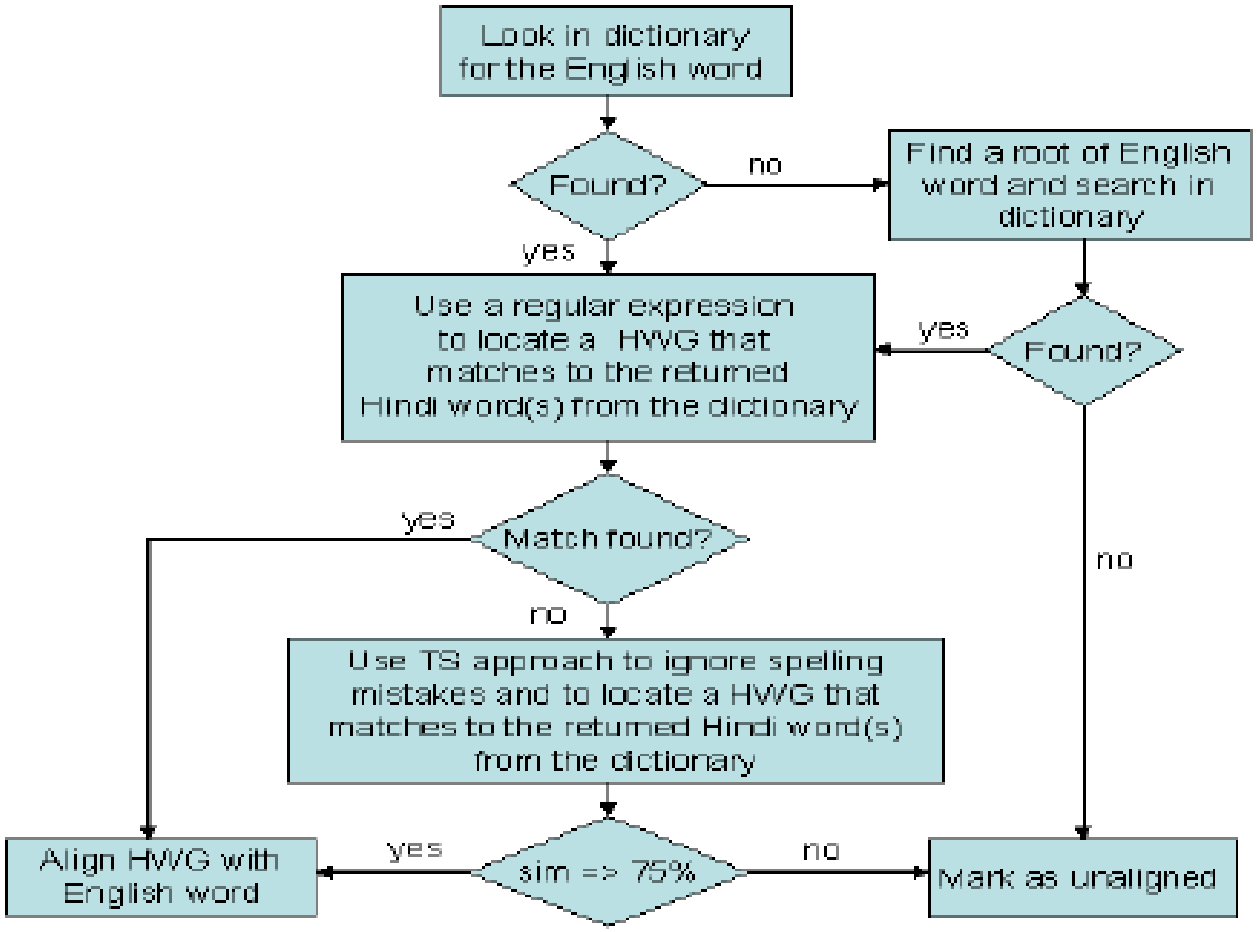

Figure 3.1 Dictionary Lookup Approach

Then nearest neighbor approach

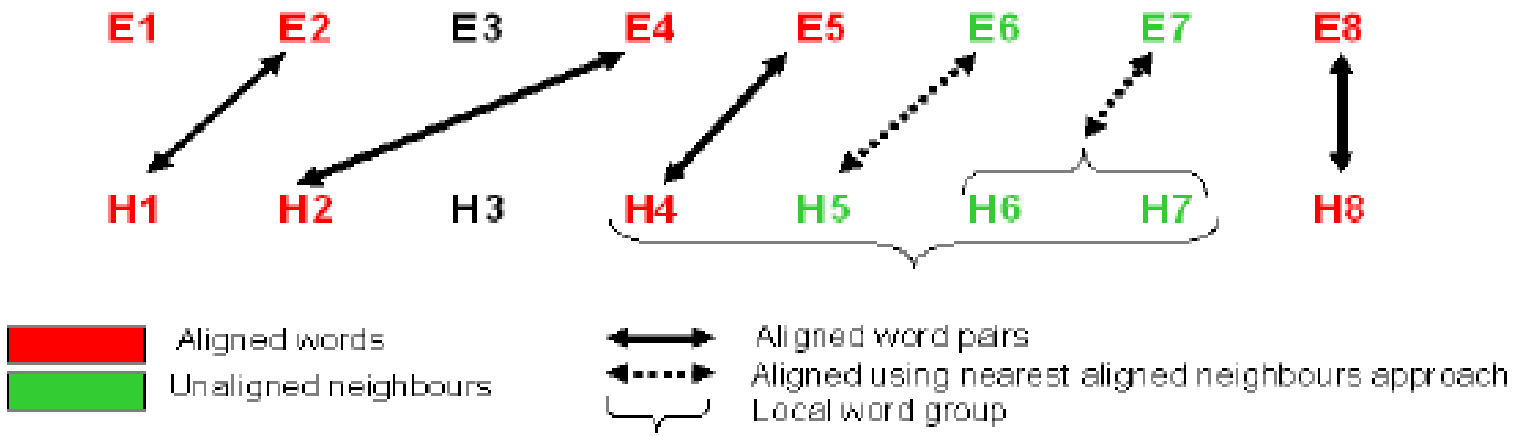

Figure 3.2 Nearest Aligned Neighbours Approach

Now lets focus on the problem:

Google Translate was used for translation.

English Sentence:

Alignment problem is difficult.

Hindi Translation:

संरेखण समस्या मुश्किल है

Also possible:

संरेखण मुश्किल समस्या है

Running Stanford Parser on it:

Pos Tags:

- Alignment/NNP is/VBZ difficult/JJ problem/NN

The parse bracketed tree:

```
(S
    (NP
        (NNP Alignment)
     )
    (VP
        (VBZ is)
         (NP
                (JJ difficult)
                 (NN problem)
        )
    )
)
```

The text Alignment

Alignment -> संरेखण

problem -> समस्या

Is -> है

Difficult -> मुश्किल

Fact: POS is same in hindi and english

POS Tags for hindi sentence:

संरेखण/NNP समस्या/NN मुश्किल/JJ है/VBZ

संरेखण/NNP मुश्किल/JJ समस्या/NN है/VBZ

Pattern Matching/Unification

The parse bracketed tree:

```
(S
    (NP
      (NNP Alignment)==(NNP संरेखण)
    )
    (VP
      (VBZ is)==(VBZ  है)                                    ----(1)
       (NP                                                  ----(2)
            (JJ difficult)==(JJ मुश्किल)
             (NN problem)==(NN समस्या)
      )
    )
)
```

Lets try to make the parse and sentence sequence consistent

Exchange: (1) and (2)

```
(S
      (NP
         (NNP Alignment)==(NNP संरेखण)
       )
      (VP
         (NP                                                    ----(2)
                  (JJ difficult)==(JJ मुश्किल)
                   (NN problem)==(NN समस्या)
          )
          (VBZ is)==(VBZ  है)                                    ----
     (1)
     )
)
```

Rule Proposed:

We can exchange nodes at the same level to make the unified tree be consistent with the Sequence of translated sentence.

Remove the English words

```
(S
    (NP
      (NNP संरेखण)
    )
    (VP
      (NP
            (JJ मुश्किल)                    -------(3)
            (NN समस्या)                     -------(4)
      )
      (VBZ  है)
    )
)
```

We get a parse tree for the second translation.

Lets make it consistent with first translation

We use the proposed rule and exchange (3) and (4)

```
(S
    (NP
        (NNP संरेखण)
     )
    (VP
        (NP
                (NN समस्या)                    -------(4)
                (JJ मुश्किल)                   -------(3)
        )
        (VBZ  है)
    )
)
```

This is a parse tree for first translation.

Both the trees can be parsed by CFG of HIndi

Here CFG

Is:

S->NP VP

VP->NP VBZ

NP->JJ NN | NN JJ

Problem: The rule is not robust

Take an example

I shot the man with ice cream

मैं आइसक्रीम के साथ आदमी को गोली मार दी

Translation itself is not correct!!!

Difficult to find alignment

के and को don’t align to anything.

What can we do?

Apply Unsupervised parsing on hindi corpus. And for scoring the parse trees we can give higher scores to the one that is consistent with parse tree generated by proposed approach.