Part-of-speech Tagging for Middle English through Alignment and Projection of Parallel Diachronic Texts

Taesun Moon and Jason Baldridge

Presenter: Yevgeni Berzak

Annotation of Historical Languages

Annotation: Marking texts written in historical languages with linguistic information.

Motivation

Diachronic Linguistics

- Language change.
- Language variation.

Case study: POS tagging for Middle English

Part of Speech Tagging

- Sequence Labeling Task: associate words in context with their syntactic categories.

“In the beginning God created the heavens and the earth.”

Part of Speech Tagging

- Sequence Labeling Task: associate words in context with their syntactic categories.

"In/PREPOSITION the/DETERMINER beginning/NOUN God/NOUN created/VERB the/DETERMINER heavens/NOUN and/CONJUNCTION the/DETERMINER earth/NOUN."

Part of Speech Tagging

- Sequence Labeling Task: associate words in context with their syntactic categories.

"In/PREPOSITION the/DETERMINER beginning/NOUN God/NOUN created/VERB the/DETERMINER heavens/NOUN and/CONJUNCTION the/DETERMINER earth/NOUN."

- Useful for syntactic parsing, morphological analysis, and many other tasks.
- Major problem – ambiguity.

Part of Speech Tagging

How to do it?

- Use statistical tagger (n-grams, ME, Transformational Tagging…)

Supervised approach:

- Use manually annotated training corpus.
- Train tagger this corpus.
- Apply tagger to new data.

Part of Speech Tagging

Middle English: 11th to 15th century

"In the bigynnyng God made of nouyt heuene and erthe."

Challenges in tagging Middle English

- Limited amount of machine readable text.
- Inconsistent orthography.
- Grammatical diversity (different genres, periods, dialects, etc..).

Part of Speech Tagging

How can we induce a tagger for Middle English?

(or any other historical language..)

Tagging a Historical Language

First approach

- Do the same as for modern languages:

  Use manually annotated data to train a tagger.

Problem:

- Very few annotated recourses for historical languages.
- Manual annotation:
  - Time, Money, Skills.
  - Error Prone

Tagging a Historical Language

- Second Approach: avoid annotation bottleneck by **Leveraging existing recourses** for relevant modern languages.
- Use parallel corpora – translations of the same text to two languages.
- Use tagging of a modern language to approximate tagging of a historical language. (Exploiting inherent similarities between the modern and the historical language)

Tagging Middle English

- Key Idea exploit parallel annotated corpora of Modern English to tag Middle English.
- Align the words
- Project the tags

In/ ? the/ ? bigynnyng/ ?...

In/PREPOSITION the/DETERMINER beginning/NOUN…

- Train a tagger on this corpus

Tagging Middle English

- Key Idea exploit parallel annotated corpora of Modern English to tag Middle English.
- Align the words
- Project the tags

In/ ? the/ ? bigynnyng/ ?...

- Train a tagger on this corpus

Tagging Middle English

- Key Idea exploit parallel annotated corpora of Modern English to tag Middle English.
- Align the words
- Project the tags

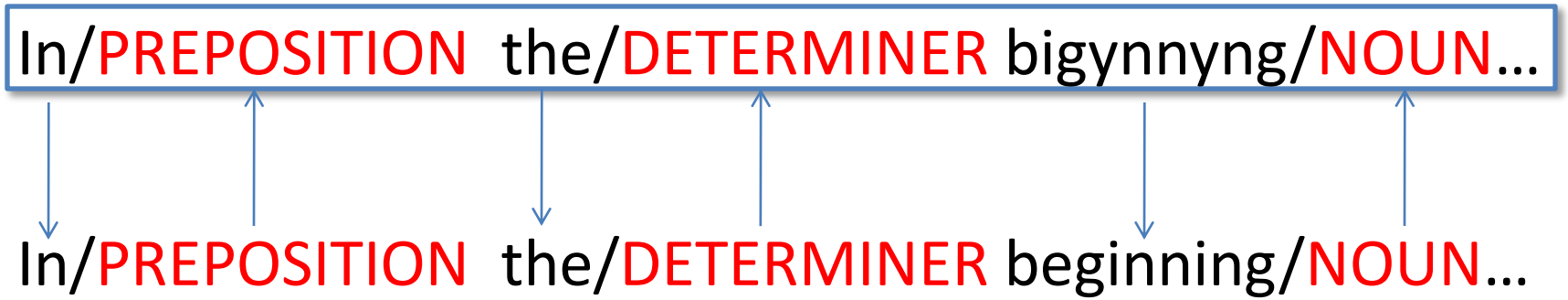
In/PREPOSITION the/DETERMINER bigynnyng/NOUN…

In/PREPOSITION the/DETERMINER beginning/NOUN…

- Train a tagger on this corpus

Tagging Middle English

- Key Idea exploit parallel annotated corpora of Modern English to tag Middle English.
- Align the words
- Project the tags



- Train a tagger on this corpus!

Tagging with Alignment & Projection

Question: Which parallel corpus can we use?

Tagging with Alignment & Projection

Question: Which parallel corpus can we use?

Answer: The Bible

- Existing (electronic) translations for many historical and modern languages.
- Relatively large - around 900,000 words.
- Clear separation of verses – facilitates sentence alignment.

Tagging with Alignment & Projection

Dice Alignment: a word in Middle English is aligned to the word in modern English that co-occurs with it most often.
To license alignment a threshold has to be passed

Giza++ Alignment: Off-the-shelf alignment Software. Uses IBM language models and HMM's.

Tagging with Alignment & Projection

Tags projection: project the majority tag of the aligned Modern English word.

Problems:

1) Alignment & projection are approximations
2) Some Middle English words are not aligned and thus don't receive tags.

Bigram Tagging

- Solution for gaps: complete missing tags with a bigram tagger.
- Bigram tagger: find the most likely tag for a word given the preceding tag.

  the/DETERMINER(t_{i-1}) bigynnyng(w_i)/NOUN(t_i)

- Training: Estimate $P(t_i|t_{i-1})$ and $P(w_i|t_i)$ from corpus counts of successfully projected sequences (Smooth unseen events).

Bigram Tagging

- Side effect: Bigram tagger for Middle English.
- Apply tagger to its training corpus.

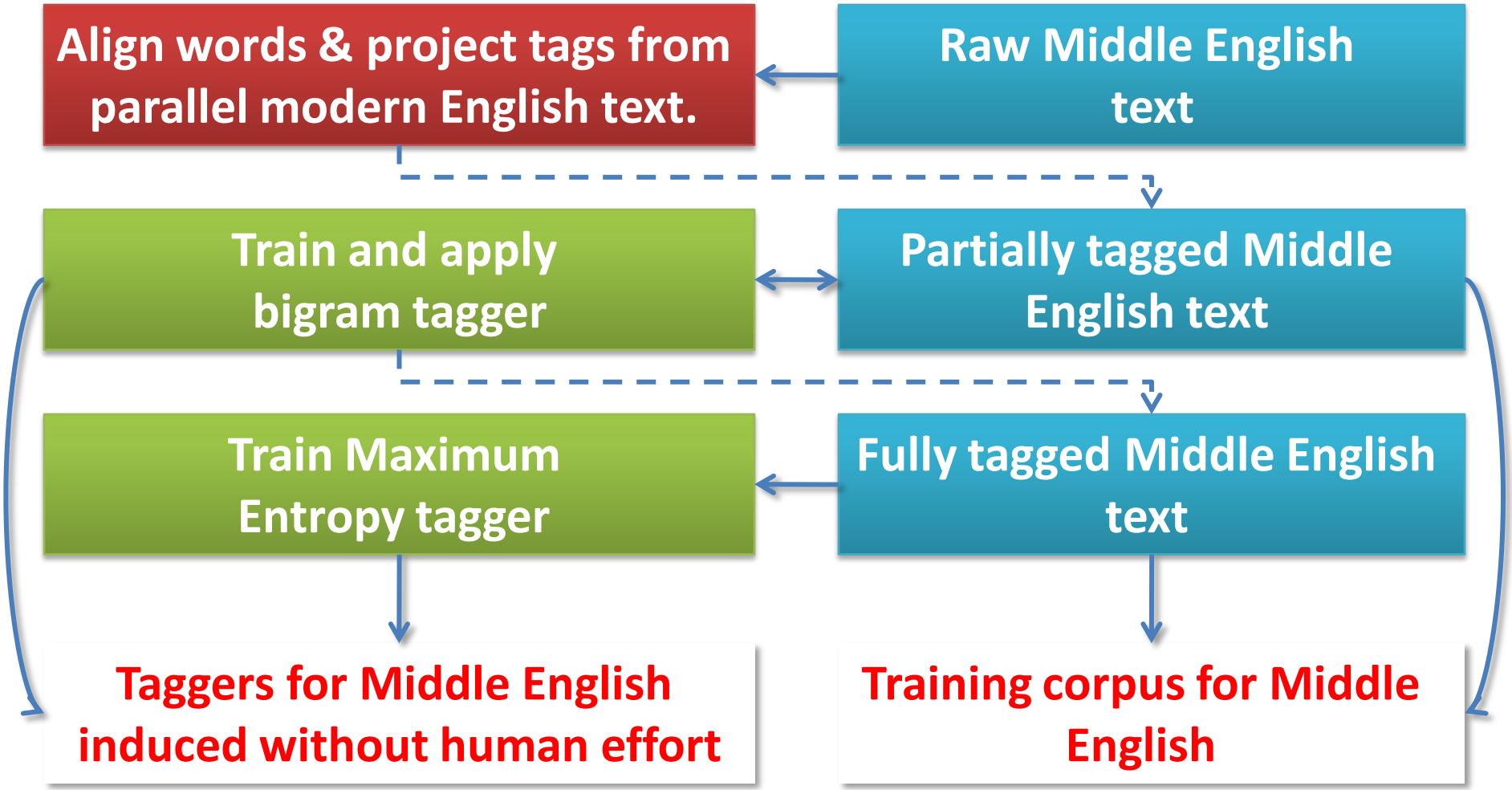
  → Retagged Middle English Bible,

  where all words have tags.

Maximum Entropy Tagging

- Use the output of the bigram tagger to train a more sophisticated tagger: C&C Maximum Entropy tagger.
- Uses many features, including two previous tags, two previous and two following words, affixes, etc…
- The induced C&C tagger can be considered as a specialized tagger for Middle English!

Recap


Align words & project tags from parallel modern English text.
Raw Middle English text
Train and apply bigram tagger
Partially tagged Middle English text
Train Maximum Entropy tagger
Fully tagged Middle English text
Taggers for Middle English induced without human effort
Training corpus for Middle English

Evaluation

- Evaluation Corpus – “Penn-Helsinki Parsed Corpus of Middle English”(PPCME).

  Tagged text samples of Middle English from 55 different sources.
- More then million words.
- Includes portions of the Bible.

Evaluation

Model	In domain (PPCME Bible)	Out of domain (PPCME other texts)
C&C trained on Modern English	56.2%-63.4%	56.2%-62.3%
C&C trained on Middle English projected tagging	**78.8%-84.1%**	**61.3%-67.8%**

- ≈20% improvement on biblical material.
- ≈5% improvement on other Middle English texts.

Discussion

- Strong domain effect.
- Performance within domain is much better, but still far from state of-the-art. Why?
- If high accuracy is needed, carefully sampled manual annotation is still a reasonable approach.
- Tagger could be used for semi-automated tagging.

To Sum Up

- A reasonably good POS tagger for historical languages can be induced with minimal human effort using alignment and projection of tags from modern languages.
- The Bible can be a useful recourse for adapting NLP tools for historical languages.
- Linguistic annotation can help us gain insight on language change and variation.