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

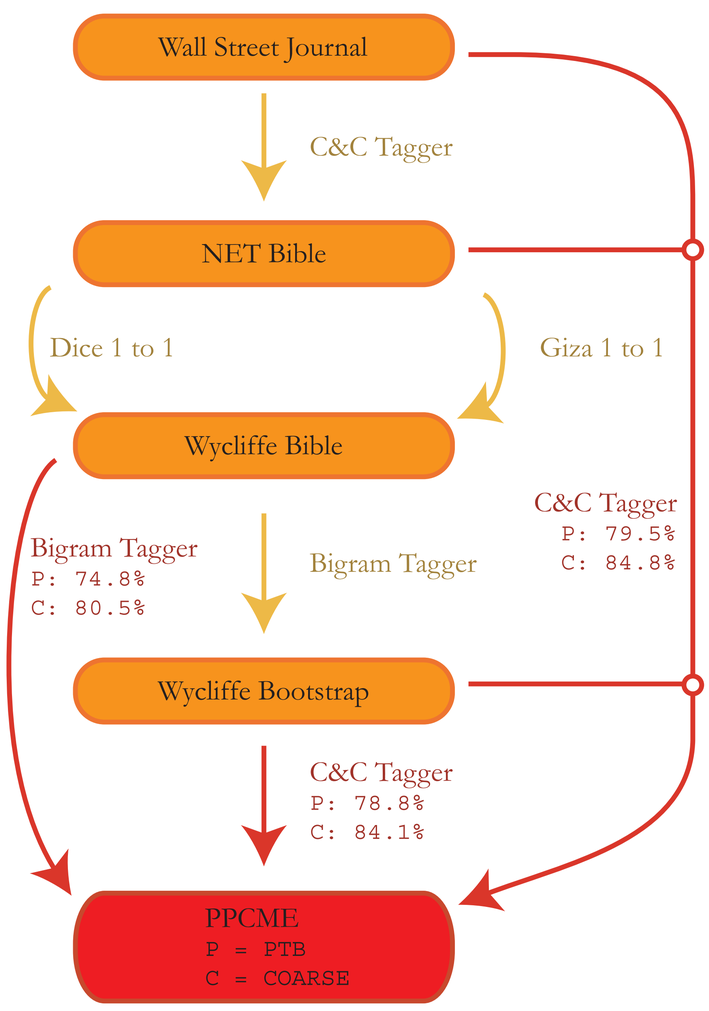
Taesun Moon and Jason Baldridge

tsmoon,jbaldrid@mail.utexas.edu

Abstract: **We demonstrate an approach for inducing a tagger for historical languages based on existing resources for their modern varieties. Tags from a Present Day English Bible are projected to a Middle English Bible using multiple alignment approaches and are smoothed with a bigram tagger. Finally, we train a maximum entropy tagger on the output of the bigram tagger on the target text and test it on tagged Middle English text.**

Approach Outline

1. Build a mapping table of words from target text to source text using a standard alignment model
2. POS tag the source text using a standard resource
3. Project POS tags from source to target
4. Train bigram tagger on all tagged resources and reapply on target
5. Train maximum entropy tagger on source and target, then reapply on target



Yellow lines represent creation of training data. Red lines represent tagger and accuracy as measured against converted PPMCE tags

Data

- The Bibles
  - Source: The New English Translation(NET) Bible (2005)
  - Target: The Wycliffe Bible (late 14th century)
- The Gold Standard: The Penn Helsinki Parsed Corpus of Middle English (PPCME)
  The PPCME [3] is a collection of Middle English texts, provided in three forms: raw, POS tagged, and parsed. It also contains portions of the Wycliffe Bible (*Genesis, Numbers, John I.1-XI.56*) making it a suitable gold standard for the present task.
- Training and testing materials
  A reduced subsection of the PPCME contemporaneous with the Wycliffe Bible was split into a training set, a test set, and the Wycliffe set.

1 In the beginning God created the heavens and the earth.
1 In the bigynnyng God made of nouyt heuene and erthe.
2 Now the earth was without shape and empty, and darkness was over the surface of the watery deep, but the Spirit of God was moving over the surface of the water.
2 Forsothe the erthe was idel and voide, and derknessis weren on the face of depthe; and the Spiryt of the Lord was borun on the watris.

The first two verses of Genesis, interlinearized. The NET Bible precedes Wycliffe's Bible.

Alignment and Projection

- Dice Coefficient [2]
  A simple, heuristic measure for creating word level alignments, it was used to create the D_1TO1 training portion for the bigram tagger.
- GIZA++ [4]
  Two separate POS tag projections from source to target were implemented through the word alignment program GIZA++. One is a direct projection from source to target based on the alignment (G_1TON). Another is by generating a mapping table of word types from source to text. In this case, the most commonly occurring POS tag for a given word type in the source was transferred to the target (G_1TO1)

Tagging

- The Curran and Clark Tagger (C&C) [1] C&C is a maximum entropy tagger, and it was used to tag the NET Bible with the Penn Treebank tagset (PTB).
- Bigram Tagger
  - Because the projection methods outlined above leave gaps in the POS tag sequence, a bigram tagger was trained on various combinations of the tagged texts to tag the target text
  - The bigram tagger trained on D_1TO1 and G_1TO1is referred to as BOOT.
- Tagsets
  Because the PPCME tagset is larger than PTB, this tagset was was mapped to PTB for evaluation purposes. A second, further reduced tagset (COARSE) was also considered.

Results

The results of the tagging attempts according to training material, tagset, and evaluation standard is presented below.

Model	Evaluate on PPCME Wycliffe		Evaluate on PPCME Test	
	PTB	COARSE	PTB	COARSE
(a) Baseline, tag NN	9.0	17.7	12.6	20.1
(b) C&C on WSJ	56.2	63.4	56.2	62.3
(c) BG on D_1TO1 + G_1TON	68.0	73.1	43.9	49.8
(d) BG on D_1TO1 + G_1TO1	74.8	80.5	58.0	63.9
(e) C&C on BOOT	78.8	84.1	61.3	67.8
(f) C&C on BOOT + WSJ + NET	79.5	84.8	61.9	68.5
(g) C&C on gold Wycliffe	n/a	n/a	71.0	76.0
(h) C&C on training	95.9	96.9	93.7	95.1

Tagging results.

Discussion

- One study [5] with a similar motivation but using a semi-automated approach achieved 96% accuracy.
- An examination learning curves and the cost of annotation revealed that our appoach is overtaken with 50 sentences for the PPCME Test and 400 sentences for PPCME Wycliffe.
- Note, however, the domain effect between items (g) and (h) in the table above. Accuracy suffers critically if C&C is trained only on the PPCME Wycliffe Bible.

References

[1] James R Curran and Stephen Clark. Investigating gis and smoothing for maximum entropy taggers. In *Proceedings of the 11th Meeting of the European Chapter of the Association for Computational Linguistics (EACL-03)*, 2003.

[2] Martin Kay and Martin Röscheisen. Text-translation alignment. *Computational Linguistics*, 19(1):121–142, 1993.

[3] Anthony Kroch and Ann Taylor. Penn-helsinki parsed corpus of middle english, second edition, 2000.

[4] Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003.

[5] David Yarowsky and Grace Ngai. Inducing multilingual pos taggers and np bracketers via robust projection across aligned corpora. In *NAACL '01: Second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies 2001*, pages 1–8, Morristown, NJ, USA, 2001. Association for Computational Linguistics.