Antecedent Prediction without a Pipeline

Sam Wiseman Alexander M. Rush Stuart M. Shieber
Harvard SEAS, Cambridge, MA

HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

Introduction

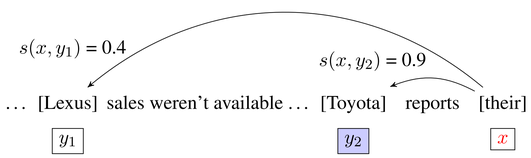
We investigate very simple antecedent prediction models that do not rely on upstream (pipelined) features (e.g., parses, named-entity tags), and instead make predictions only using words and sentence boundaries. We attempt to understand where un-pipelined models go wrong, and how they might be improved.

Why Go Un-pipelined?

- Much simpler
- No error propagation/accumulation
- More applicable to low-resource/social-media settings where upstream features are less reliable
- Congenial to "NLP from Scratch" view (Collobert & Weston, 2011)

Task/Approach

- Ranking of antecedents for anaphoric mentions



- Models use only word and distance-based information

Very Important Caveat

Though our ranking models use only word and distance information, mentions are automatically extracted **using pipelined parse information** (as is common).
Accordingly, this work **should be interpreted as a thought-experiment/attempt to get an upper bound** on how un-pipelined models might perform.

Models

- Ranking function uses a simple MLP
- The MLP consumes distance embeddings, as well as word-level information from both current mention and antecedent
- Word-level information consists of embeddings of words in mention, and embeddings of fixed word-windows before and after mention
- Embeddings are aggregated with either:
  - Max-Pooling
  - Convolution, followed by max-over-time (Kim, 2014)
  - LSTM (Hochreiter & Schmidhuber, 1997)

Experiments

- Experiments on CoNLL 2012 English Development set
- We compare with antecedent-ranking MLP of Wiseman et al. (2015), which uses pipelined features

Results

Model	Acc.
Wiseman et al. (2015)	82.58
Max-Over-Time	70.92
Convolution	72.65
LSTM	77.40

Table 1: Antecedent prediction accuracy of models and baseline on CoNLL Development set.

Error Analysis

	Errors		
	HM	No HM	Pron.
Wiseman et al. (2015)	588	522	1146
Max-Over-Time Model	1513	608	1646
Convolutional Model	1358	607	1577
LSTM Model	1028	537	1362
Total Mentions	4677	973	7302

Table 2: Mentions are partitioned column-wise as nominal or proper with (previous) head match (HM), nominal or proper with no previous head match (No HM), and pronominal.

Discussion

- **Hypothesis: un-pipelined models are bad at head-finding**
- For nominal/proper errors, predicted antecedents are semantically reasonable, but model seems to be ignoring heads:
  - For mentions with a previous head-match where the Convolutional model erroneously disagrees with the pipelined MLP, $\approx 84\%$ involve the Convolutional model predicting *an antecedent with a different head*
  - LSTM only improves on these sorts of errors by $\approx 6\%$

Mention	(A) True Antecedent	Predicted Antecedent
the Straits [Foundation]	the Straits [Foundation]	the Straits [Association]
those Jewish [sacrifices]	the [sacrifices]	the [people] of Israel
the [water]	[water]	their sinking fishing [boat]

Table 3: Example mentions, which the baseline MLP correctly predicts (middle column), but the Convolutional Model (right column) does not. Heads of each mention (unseen by the Convolutional Model) are in brackets.

Further Evidence

- Performance of word-only models decreases as mentions get longer!
- Plausibly because head-finding more difficult in such cases

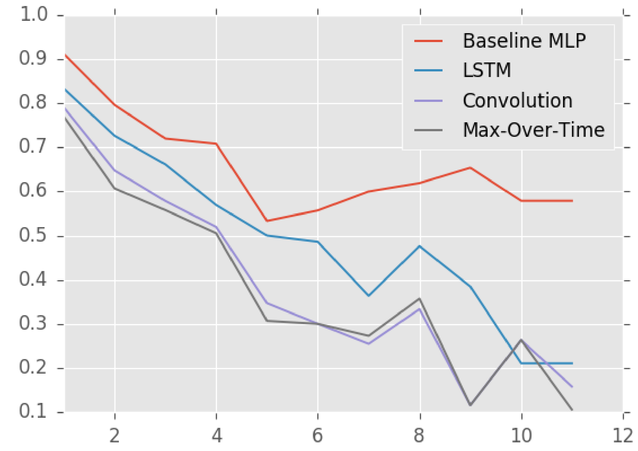


Figure 1: Percentage of antecedents in the CoNLL 2012 development set predicted correctly, by mention length.

Pronominal Errors

- Discrepancy between word-only and pipelined model on pronouns less easy to diagnose:
- Errors tend to involve un-pipelined models either predicting number- or gender-incompatible antecedents, or non-pronominal antecedents when a different pronominal antecedent will do

Future Directions

- Models somewhat more sensitive to syntax may help!
- A natural opportunity for attention-based modeling!

{swiseman,srush,shieber}@seas.harvard.edu