Memory-based named entity recognition in tweets

Antal van den Bosch
Centre for Language Studies, Radboud University Nijmegen, The Netherlands

Toine Bogers
Royal School for Library and Information Science, Copenhagen, Denmark

Making Sense of Microposts 2013 Concept Recognition Challenge

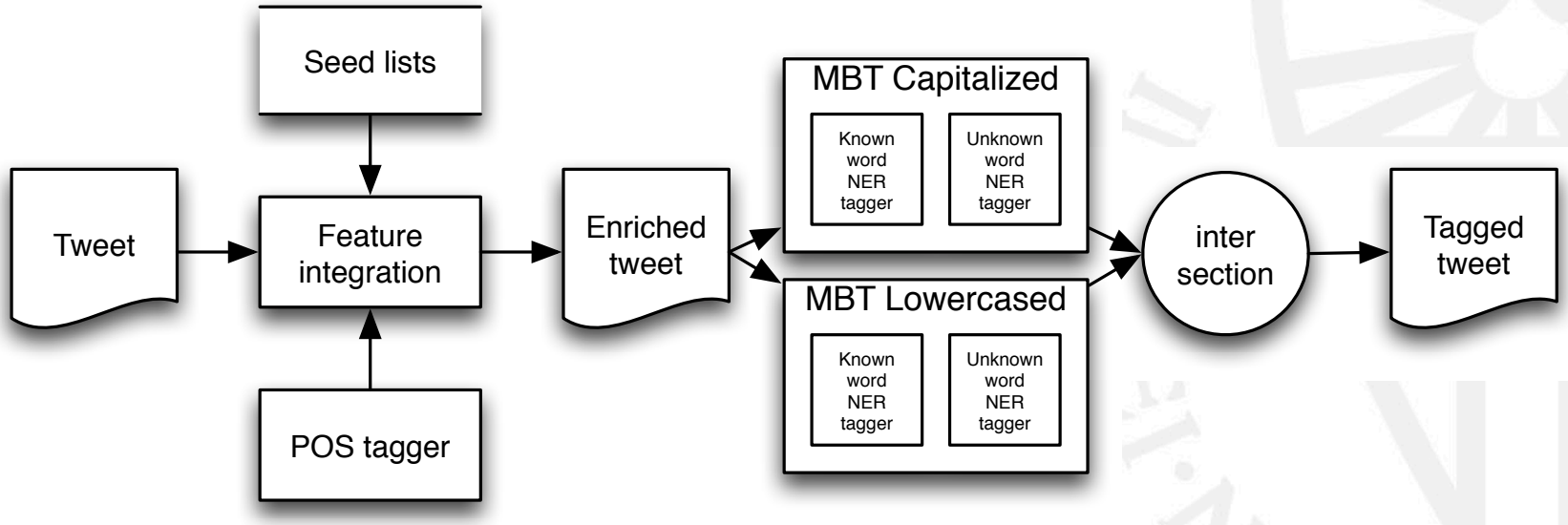
Memory-based named entity recognition in tweets

- Memory-based tagging appled to NER
  - Token-level 'IOB' tagging based on k-NN classification
    - All entity words tagged with I-<entity_type>
    - Except first word of an entity immediately preceded by another entity: B-<entity_type>
    - All non-entity words tagged with O
  - Sequence processor operating from left to right, taking into account previous decisions
  - Separate models for known and unknown words

- Two taggers
  - *Capitalized:* Tagger trained on original data 'as is', with capitalization
  - *Lowercased*: Tagger trained on lowercased data
  - Twitter data is inconsistent in casing (and in other spelling/formatting)

- Fusing the output of *Capitalized* and *Lowercased*
  - By taking intersection: only named entities detected by both taggers

System Architecture



Workflow

1. A new incoming tweet is first enriched by **seed list** information;
2. The tweet is also **part-of-speech tagged**
3. The enriched tweet is then processed by **two MBT taggers**:
   a. The first tagger is trained on the original training data with all capitalization information intact;
   b. The second tagger is trained on a lowercased version of the training set.
4. The taggers both assign **IOB-tags** to the tokens constituting named-entity chunks;
5. The two MBT modules generate partly overlapping predictions. Only the named entity chunks that are fully identical in the output of the two modules, i.e. their **intersection**, are kept;

6. The result is a tweet annotated with named entity chunks.

Resources

Named entity recognition **training data**:

- Official (version 1.5) training data provided for the challenge;
- Training and testing data of the CoNLL-2003 Shared Task;
- Named-entity annotations in the ACE-2004 and ACE-2005 tasks;
  - http://projects.ldc.upenn.edu/ace/
- For all tokens, add a POS tag (Penn Treebank).

Gazetteer/seed lists:

- 8.4 million geographical names taken from Geonames.org
  - http://download.geonames.org/export/dump/allCountries.zip
- 499 thousand person names and 40 thousand organization names from the JRC Names corpus
  - http://optima.jrc.it/data/entities.gzip
- For all tokens in training and test data, add a code (G, P, O, GP, PO, GO , GPO) if it occurs in one or more seed lists.

Results (1)

- On self-made development split of challenge data
  - Development set: 22,358 tokens, 1,131 named entities
- Intersection of capitalized and lowercased models boosts precision:

Table 1: Overall named entity recognition scores by the system and its components

Component	Precision	Recall	F-score
Capitalized	54.62	63.75	58.83
Lowercased	57.38	62.86	60.00
Intersection	65.82	57.21	61.21

Results (2)

- If gazetteer features are disabled,
  - overall precision increases slightly from 65.8 to 66.1, but
  - recall decreases from 57.2 to 54.9,
  - Leading to a lower overall F-score of 60.0
- Person names >> Location names > Organization names >> Misc names

Table 2: Overall named entity recognition scores on the four entity types

Named entity type	Precision	Recall	F-score
Person	75.90	69.52	72.57
Location	54.95	44.25	49.02
Organization	47.46	39.25	42.97
Miscellaneous	17.54	11.39	13.85

Thank you

a.vandenbosch@let.ru.nl
tb@iva.dk