Off to a Cold Start: New York University's 2013 Knowledge Base Population Systems

Ralph Grishman
Computer Science Department
New York University
New York, NY 10003 USA
grishman@cs.nyu.edu

1 Slot Filling

New York University submitted two runs this year: a KBP Slot Filling run and a Cold Start run.

The Slot Filling run used the same system as last year (Min et al. 2012), with only the minimal changes needed to address the revised specifications. These included changes in the provenance information and collapsing employee_of and member_of slots. We included the confidence estimation step (Li and Grishman 2013) which had been implemented last year after the official run and was described in last year's proceedings (Min et al. 2012).

2 Cold Start

Our goal for Cold Start was to become familiar with the task and to understand the importance of various features, as we had in past years for Slot Filling.

We began by building, over the course of 11 days, a bare-bones system to perform the task. We drew upon existing components from the NYU Jet system and components previously developed for KBP Slot Filling. The only major component imported from elsewhere was the Tratz-Hovy dependency parser from ISI (Tratz and Hovy 2011).[1]

The Cold Start processing is performed in two phases. The first phase analyzes each document in isolation and is thus easily parallelized. The output of the first stage is an APF file, the format adopted by the ACE [Automatic Content Extraction] program. The APF file records the entities, values, and binary relations extracted from the document. One APF file is generated for each source document. The second phase reads in all the documents and their APF files, links entities across documents, and writes a knowledge base.

The first phase involves the following steps:

- **Dictionary look- up.**
- **Name tagging**. We initially used an MEMM [Maximum Entropy Markov Model] trained on ACE data.
- **TIMEX2 tagging**.
- **Part-of-speech tagging**.
- **Dependency parsing**. Using the Tratz-Hovy parser.
- **Reference resolution**. Using hand-coded rules.
- **Entity extraction**. Component taken from our ACE system.
- **Relation extraction**, using both sequential and dependency patterns, taken from our Slot Filling system. Combines hand-coded rules and rules produced by bootstrapping combined with manual review. (Because of limited development time and a desire to keep the initial implementation simple, we did not include the relation extractor based on

[1] This parser was selected for its speed; it uses an *easy-first* strategy, and produces labeled dependencies at roughly 75 sentences per second.

distant supervision which we have successfully used for Slot Filling.)

- **APF generation**.

The first phase requires about one core-day to process the 50,000 documents of the 2013 Cold Start corpus (in practice we ran for 8 hours on 3 cores). This represents roughly 30 documents per minute per core.

The system operated on the xml-format files provided. No attempt was made to clean up the input by removing short text fragments.[2]

The second phase reads in all documents and operates entirely in memory. It first takes each entity in each document which includes a named entity mention and maps that document-level entity to a global entity. This is done by a simple deterministic rule for name mapping (see Appendix); no attempt is made to do more sophisticated entity coreference. We then iterate over the set of global entities and write out all the knowledge-base information for that entity: the mentions, the canonical mentions, and the slot-filler triples.

For single valued slots we adopted the same strategy used in Slot Filling: a value attested by multiple documents was preferred over one obtained from only a single document (more likely a typo or error); among values attested by multiple documents, we preferred the entity with the longer name.

The knowledge base which constituted our official submission was 3843106 lines, including 108771 person entities, 40669 org entities, and 16311 GPEs.

3 Scoring

The system we developed was parameterized so that it could handle both 2012 and 2013 Cold Start tasks, which differed slightly in slots and knowledge base format. Because no scorer was available in time for the needed analysis, we also developed a scorer able to handle both tasks (which used different assessment file formats). (This later became the official scorer for this task.) This scorer operates based on exact string match; document/offset information is not used. This seemed most appropriate for supporting continued development after the formal submission. Responses which do not appear in the assessment file are ignored; they are not counted as correct or incorrect.

Because of the small pool involved in Cold Start (3 teams including ourselves), many new responses produced during subsequent system development would not be part of the official assessment, making it difficult to properly judge system revisions. We therefore added a simple interactive tool to the scorer to manually assess all new responses.

4 Cold Start > Slot Filling + Entity Linking

The Slot Filling and Entity Linking tasks were initially created to decompose the more general task of knowledge base creation from text. So in simplest terms Cold Start is putting these two back together. However, other aspects of the evaluation – in particular the variability and surprise nature of the test corpus – raise important additional challenges regarding the ability to handle previously unseen sources. In this section we briefly illustrate some of these issues.

4.1 Slot Type Distribution

The Slot Filling evaluation uses a large corpus including a large portion of national and international news, and only changes modestly from year to year. This provides some stability in terms of the slots which predominate for the task (Tables 1 and 2). NYU has taken advantage of this stability and the skewed distribution of slot fills to get good Slot Filling performance over several years from a focus on the patterns for a small number of slots.

Cold start provides greater discretion for the evaluators, first in picking the domain/corpus and then in selecting the queries. The domain heavily influences the information available, and this effect can be magnified by the choice of queries. In 2012 the domain was the University of Pennsylvania; in 2013, Kentucky.

As is true in Slot Filling, the distribution across slots is highly skewed; the 5 top slots constitute more than 50% of the relations involved in queries.[3] Not surprisingly, in the university

[2] Names with three or more consecutive blanks were rejected, thus blocking names from spanning text blocks.

[3] Slot counts were based on the number of times the slot was part of a response to a query that was assessed as correct. For two-hop queries, we counted both hops for each response.

domain org:students and per:schools_attended (reciprocals of the same relation) dominated, together accounting for over a quarter of the slots (Table 3). This year's local news provided a more 'normal' distribution, though we note that participants in local news seem to be identified more often by their town (gpe:residents_of_city), and less by their title, than is common in national and international news (Table 4).

This skewed slot distribution increases the volatility of the scores with respect to changes in the extraction system. This volatility is magnified by the fact that particular text sources will use particular locutions to express a given relationship. To study this effect, we created a text corpus consisting of the provenance of all responses assessed as correct by LDC, a total of 1788 items for the 2013 evaluation. Of these, 293 were instances of the gpe:residents_of_city relation which contained the sequence ", of". This sequence was used for

person, of *city*

or

person, *age*, of *city*

(e.g., "Fred Smith, of London" or "Fred Smith, 66, of London"). Thus essentially one-sixth of the 0-hop score depended on recognizing this one construct.[4] This pattern is in particular the style of one of the news sources, *the Times-Tribune*.

To a lesser extent this same effect was observed for the Penn domain; there were over 100 instances of the expression "received his/her PhD. from/at" to convey the students / schools attended relation.

Table 1. Top slots for 2012 Slot Filling evaluation

slot	# of equivalence classes
per:title	236
org:top_members_employees	177
per:member_of	107
per:children	100
org:alternate_names	96
per:employee_of	72
per:cities_of_residence	62
org:subsidiaries	59

Table 2. Top slots for 2013 Slot Filling evaluation

slot	# of equivalence classes
per:title	230
org:top_members_employees	130
per:employee_or_member_of	125
org:alternate_names	89
per:alternate_names	63
per:children	57
per:cities_of_residence	53
per:age	51

Table 3. Top slots for 2012 Cold Start evaluation

slot	frequency in responses
org:students	441 (18.16%)
org:employees	407 (16.76%)
per:title	337 (13.88%)
per:employee_of	223 (9.18%)
per:schools_attended	173 (7.13%)
org:top_members_employees	101 (4.16%)
org:membership	97 (4.00%)

Table 4. Top slots for 2013 Cold Start evaluation

slot	frequency in responses
gpe:residents_of_city	693 (15.65%)
gpe:employees_or_members	570 (12.87%)
org:employees_or_members	555 (12.53%)
per:title	425 (9.60%)
pre:employee_or_member_of	199 (4.49%)
gpe:headquarters_in_city	196 (4.43%)
org:top_members_employees	155 (3.50%)

4.2 Influence of name tagging

We found Cold Start to be considerably more sensitive to the quality of name tagging than Slot Filling had been. For Slot Filling, the system is given the names being queried in advance, including extent and type information. In contrast, none of this information is given in advance for the Cold Start task; a name extent or type error involving one of the queried names often leads to a failure to respond to the query.

We initially used a MEMM name tagger that had provided good performance in prior tasks

This meant that if the same first hop participated in several two-hop responses, it was counted multiple times. This seemed appropriate in recognizing the relative importance of the first hop in the overall score.

[4] In contrast, this pattern did not contribute any slot fills for the 2012 evaluation.

when training and test corpora were generally similar, but proved less than satisfactory here. After trying several taggers, we used a simple HMM tagger with six states per name type, trained on ACE data, which performed better on novel data sources. This produced only a small score improvement (Table 5) using the 0-hop metric and the officially assessed data, but inspection of the new unassessed data indicated consistent gains.[5] A minor change was made to name coreference, but by far the largest gain (nearly doubling the 0-hop score) was achieved by adding the single "*person*, of *city*" pattern mentioned earlier. To better assess the impact of this small change, we hand-assessed all the unassessed responses at this point and got a 0-hop score of 37.5. (This exaggerates the gain, since only assessments related to our system improvements were added, but it shows the need to augment the assessments in order to accurately gauge overall system performance.)

Table 5. 0-hop scores

system	P	R	F1
official submission (MEMM name tagger)	45.8	7.3	12.6
HMM name tagger	52.0	7.4	13.0
constraint on name coreference	52.7	7.3	12.9
added *person*, of *city* pattern	62.8	15.3	24.6
augmented assessments	71.5	25.4	37.5

With the added pattern but only official assessments, the combined F1 score (0-hop + 1-hop scores, corresponding to the 2012 official metric) was 16.8%.

Acknowledgements

I would like to thank Bonan Min for his assistance in preparing the new KBP Slot Filling data and Xiang Li for his work on confidence measures.

Appendix

Name mapping rules for coreference:

- remove corporate suffixes (Co., Corp., etc.)
- remove initials (A., B., etc.)
- replace all whitespace sequences and hyphens by _
- delete all characters except letters and _
- reduce to lower case
- append entity type (per, org, or gpe)

Names which yield the same string under this mapping are treated as coreferential. Names which contained a character other than a letter, hyphen, comma, period, semicolon, ampersand, or single quote were ignored.

References

Xiang Li and Ralph Grishman 2013. Confidence estimation for Knowledge Base Population. In *Proceedings of Recent Advances in Natural Language Processing,* Bulgaria.

Bonan Min, Xiang Li, Ralph Grishman, and Ang Sun. New York University 2012 System for KBP Slot Filling. *Proceedings of the 2012 Text Analysis Conference.*

Stephen Tratz and Eduard Hovy. 2011. A Fast, Accurate, Non-Projective, Semantically-Enriched Parser. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Edinburgh, Scotland

[5] The number of query names which were missed dropped from 37% with the MEMM tagger to 27% with the HMM tagger,