UFV in the TAC 2015 Cold Start Knowledge Base Population Track

David Johnson
U. Fraser Valley
Abbotsford, BC, Canada
david.johnson3@student.ufv.ca

Gabriel Murray*
U. Fraser Valley
Abbotsford, BC, Canada
gabriel.murray@ufv.ca

Benjamin Tremblay†
U. Waterloo
Waterloo, ON, Canada
b3tremblay@uwaterloo.ca

Jeremy Wright
U. Fraser Valley
Abbotsford, BC, Canada
jeremy.wright@student.ufv.ca

Abstract

We briefly describe our system for the TAC 2015 Cold Start Knowledge Base Population (KBP) track.

1 Introduction

University of the Fraser Valley (UFV) participated in the Cold Start KBP task for the first time in 2015. The system was developed from scratch by undergraduate students during the Summer of 2015. In this paper, we describe the entity clustering and slot-filling components, and also describe some of the issues faced during development.

2 Entity Clustering

Entity mentions are identified using the NLTK chunk package. For linking, these mentions are first sorted by length. Then, a greedy loop is run to group mentions until there are no mentions left ungrouped. On each iteration of the loop, it compares all currently ungrouped mentions to the longest ungrouped mention via a string similarity metric, described below. Each mention which scores above some threshold (an algorithm parameter) is grouped with the longest mention, and then the loop repeats. When this process is finished, the groups correspond to entities (any two mentions in the same group are linked, and the groups partition the mentions).

Our system uses the Jaro-Winkler metric, which in turn is an extension of the Jaro metric. The Jaro string metric (Jaro, 1989) is defined as

$$\Phi_J(s_1, s_2) = W_1 \cdot \frac{c}{L_1} + W_2 \cdot \frac{c}{L_2} + W_t \cdot \frac{c - \tau}{c}$$

where W_1 is the weight given to the first string, W_2 is the weight given to the second string, W_3 is the weight given to transpositions, c is the number of characters in common between the two strings, L_1 and L_2 are the length of the two strings, respectively, and τ is the number of characters that are transposed. In these experiments, the weights were all set to 1/3.

The Jaro-Winkler metric (Winkler, 1990; Herzog et al., 2007) extends the Jaro metric as follows:

$$\Phi_w(s_1, s_2) = \Phi_J(s_1, s_2) + i \cdot 0.1 \cdot (1 - \Phi_J(s_1, s_2))$$

The string similarity measure is computed by comparing the tokens of each mention via the Jaro-Winkler metric, and taking the mean of the best possible score over all possible combinations. For example, the two mentions Barack Obama and Obama, Barack Hussein would obtain a perfect score, as both mentions have the tokens Barack and Obama. Unmatched tokens are not penalized.

In our evaluations on TAC 2014 data, this greedy loop approach outperformed more sophisticated methods such as hierarchical clustering and the generative approach described by Doan et. al (2012). Specifically, it achieved 85% entity accuracy and 92% cluster accuracy, meaning that (on average) 85% of mentions of any given entity are in the same cluster, and 92% of mentions in any given cluster refer to the same entity.

*Primary contact author
†Formerly at UFV

The entity component did not make use of any external resources such as Wikipedia. Our entity clustering code is freely available[1].

3 Slot-Filling

We trained a Logistic Regression classifier on Stanford's annotated dataset of relations[2]. The features used for relation detection include lexical (bag-of-words) features as well as part-of-speech features, for both the entity spans as well as the intervening text spans. When testing the system using cross validation on the Stanford annotated dataset the system achieved the following scores:

Precision	Recall	F-Score
0.416	0.354	0.382

The slot-filling component also did not make use of any external resources or knowledge bases such as Wikipedia.

4 Interactive KB Population

Our original goal was to design an interactive system to assist manual creation of a KB. The idea is that the system would suggest entity matches and relations, and allow for human editing and manual KB population. The interactive system was not finished in time for the TAC submission, but a version of the prototype tool is described by Wright et. al (2016). We plan to continue development of the UI for use in the TAC 2016 task.

5 Issues

We were still developing the system as the submission deadline approached, and so the evaluation document set was not fully processed. Some documents were not analyzed at all, and others were severely truncated, e.g. only the first paragraph was extracted, in order to submit on time. In addition, there was a systematic error in the calculated offsets. For that reason, our 2015 scores were essentially zero and we report results above on the similar 2014 data.

During the development and post-mortem stages, it also became clear that the NLTK entity detection component is not sufficient for this data. In particular, it did not deal well with recognizing multi-token entities. We are currently working on an in-house entity recognition and classification system.

6 Conclusion

We have briefly described our entity clustering and slot-filling components, and reported results for the TAC 2014 data and the Stanford relations dataset. We plan to participate in the 2016 Cold Start KB task with a fully interactive system.

[1] https://github.com/bhstremblay

[2] http://nlp.stanford.edu/software/mimlre.shtml

References

AnHai Doan, Alon Halevy, and Zachary Ives. 2012. *Principles of data integration*. Elsevier.

Thomas N Herzog, Fritz J Scheuren, and William E Winkler. 2007. *Data quality and record linkage techniques*. Springer Science & Business Media.

Matthew A Jaro. 1989. Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida. *Journal of the American Statistical Association*, 84(406):414–420.

William E Winkler. 1990. String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage.

Jeremy Wright, Gabriel Murray, and Ben Hachey. 2016. An interface for assisted curation of knowledge bases from unstructured text. In *Proc. of HICSS 2016, Kauai, USA*.