Volume 7, Issue 5, May 2017 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Augmenting Knowledge Base with Information Extraction

Vaishali Bhargava
M.Tech Scholar, Department of Computer Science Engg.,
AKTU, Uttar Pradesh, India

Sapna Singh
Faculty, Department of Computer Science& Engg.,
AKTU, Uttar Pradesh, India

Abstract-Occupying Knowledge Base with new knowledge facts from reliable text resources usually contains linking name mentions to entities and identifying relationship between entity pairs. The main goal of KBP is to circulate research in discovering facts about entities and expanding a structured knowledge base with this information . The task often suffers from errors propagating from upstream entity linkers to downstream relation extractors. We get optimal assignment by addressing the coherence among preliminary local predictions: whether the types of entities meet the expectations of relations explicitly or implicitly, and whether the local predictions are globally compatible. There are two new tasks - Cross-lingual Entity Linking: Given a set of multi-lingual queries, the system is required to provide the ID of the English KB entry to which each query refers and cluster NIL queries without KB references; and Temporal Slot Filling: given an entity query, the system is required to discover the start and end dates for any identified slot fill.

Keywords-personal knowledge graph population, personal assertion detection, relation detection, slot filling, user profiling for content recommendation, rich annotation guided learning, task definition

I. INTRODUCTION

Recent advances in NLP have made it possible to build structured KBs from online resources, at an unique scale and much more efficiently than traditional manual edit. However, in those KBs, entities which are popular to the community usually contain more knowledge facts. While most other entities often have fewer facts and knowledge facts should be updated as the development of entities, such as changes in the cabinet, a marriage event, or an acquisition between two companies, etc. The main goal of the Knowledge Base Population track at Text Analysis Conference is to gather information about an entity that is scattered among the documents of a large collection, and then use the extracted information to populate an existing knowledge base (KB). KBP is done through two separate sub-tasks - Entity Linking and Slot Filling. For both tasks, the system is given a query contains a name and a document in which this name appears. For Slot Filling, the system must determine from a large source collection of documents the values of specified attributes of the entity, such as the age and birthplace of a person or the top employees of a corporation.

To summarize, the following improvements-

A. Defined a new task, Cross-lingual Entity Linking, and prepared its annotation guideline and training corpora;

B. Defined a new task, Temporal Slot Filling, and prepared its annotation guideline and training corpora; • Added clustering of entity mentions without KB entries into the Entity Linking task, and developed a new scoring metric incorporating NIL clustering;

C. Made systematic corrections to the slot filling guidelines and data annotation;

D. Defined a new task, Cross-lingual Slot Filling, and prepared its annotation guideline, in anticipation of future evaluations.

The experiments on a real-world case study show that this framework can eliminate error propagations in by taking relations' argument type expectations and global compatibilities into account, thus outperforms the approaches based on state of the art relation extractors by a large margin.

II. TASK DEFINITION AND EVALUATION METRICS

It summarizes the tasks conducted at KBP 2011. More details regarding data format and scoring software can be found in the KBP 2011. The overall goal of KBP is to automatically identify salient and novel entities from multiple languages, link them to corresponding Knowledge Base (KB) in a target language, then discover attributes about the entities and finally expand the KB with any new attributes. A source language S and a target language T, Figure 1 represents the general architecture of current KBP tasks.

It has been estimated that one of every fifty lines of database application code involves a date or time value. In fact, many statements in text are temporally qualified. For example, most of the slot types change over time and thus can be temporally bounded. Temporal Information Extraction is also of significant interest for a variety of NLP applications such as Textual Inference, Multi-document Text summarization and Template Based Question Answering .

III. DATA ANNOTATION

The description of the data annotation for KBP are presented in a separate paper by the Linguistic Data Consortium. The English text corpus is unchanged from KBP2010, consisting of 1,286,609 newswire documents, 490,596 web documents. A manual annotation for the 2010 slot-filling task was included along with the pooled system outputs and the pooled slot fills were then manually evaluated; the assessors did not know which fills came from the manual annotation. When the manual annotation was then scored against the assessments. One-third of the manual fills were considered incorrect by the assessors. Some errors could be attributed to revisions in the guidelines in preparation for assessment, but more generally the low correctness reflected under specification of the slot fill guidelines, particularly for some of the most common slots.

IV. MONO-LINGUAL ENTITY LINKING

A. General Architecture

A typical KBP mono-lingual entity linking system architecture is encapsulated . It contains these steps:

1) query expansion - expand the query into a richer set of forms using Wikipedia structure mining or co reference resolution in the background document;
2) candidate generation - finding all feasible KB entries that a query might link to;
3) candidate ranking - rank the probabilities of all candidates;
4) NIL detection and clustering - detect the NILs which got low confidence at matching the top KB entries from step (3), and group the NIL queries into clusters.

B. Ranking Features

Query expansion techniques are alike across systems, and KB node candidate generation methods normally achieve more than 95% recall. Even after we introduced the new NIL clustering component in this year's evaluation, systems achieved very high performance in clustering itself. Therefore, the most critical step is ranking the KB candidates and selecting the best node. It's encouraging to see many new and interesting ranking features have been invented during each year's evaluation.

C. Performance

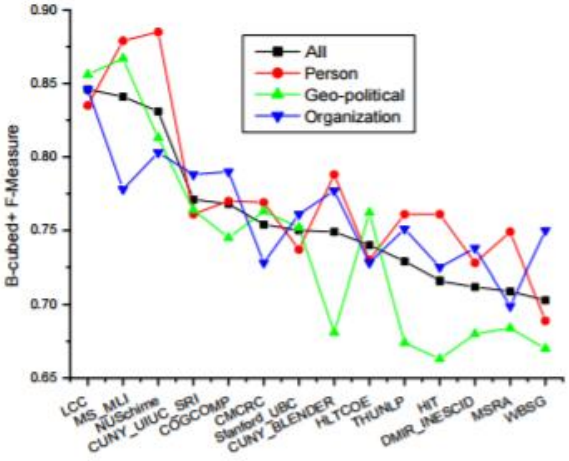
There are two principal challenges of entity linking: the same entity can be referred to by more than one name string and the same name string can refer to more than one entity. From table we can roughly estimate that KBP2011 includes more ambiguous Non-NIL entities than KBP2010.

Difficulty	Year	All	NIL	Non-NIL
Ambiguity	2010	12.9	9.3	5.7
	2011	13.1	7.1	12.1
Variety	2010	2.1	1.7	2.5
	2011	1.6	0.9	2.4

However, comparing the performance on the same KBP2010 data set, we can see almost all systems achieved significant improvement in 2011.

D. Performance of Various Entity Types

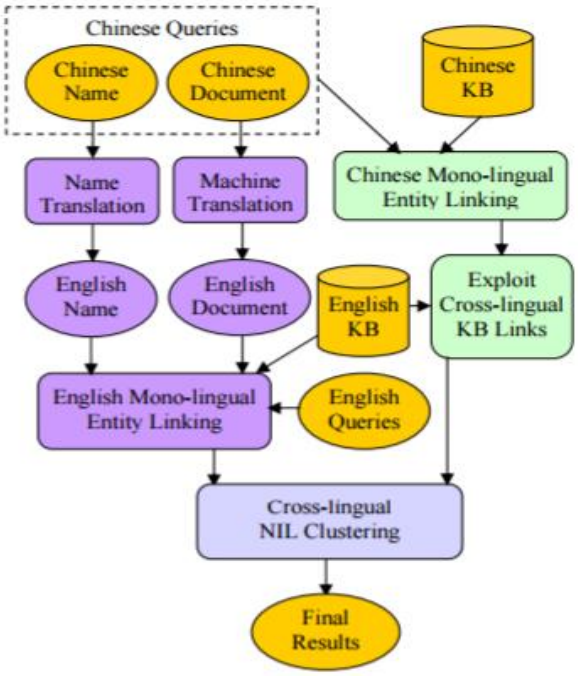
Figure shows the F-measure scores of the top 14 systems on various entity types. We can see that systems generally performed the best on person entities, and the worst on geo-political entities. However the rank of overall performance is not consistent with the rank of individual types.



V. CROSS-LINGUAL ENTITY LINKING

A. General Architecture

There are two basic approaches to cross-lingual entity linking as illustrated in Figure



- Pipeline A: Translate a Chinese query and its associated document into English, and then run English mono-lingual entity linking to link the translated query and document to English KB system.
- Pipeline B (Chinese Entity Linking + Cross-lingual KB linkages): Apply Chinese Entity Linking to link a Chinese query to Chinese KB, and then use cross-lingual KB linkages to map the Chinese KB node to English KB node and HITS system that used external hyperlinks, image similarity and templates. It's hard to tell which pipeline is better. Each method has its own limitations in terms of quality and portability. .

VI. CONCLUSION

We examined that cross-lingual training data contains higher percentage of equivocal names than mono-lingual data. we notice a few annotation errors by checking some queries randomly. If time and funding permit in KBP2012, we should check human annotation performance and do better annotation quality control for this task. It is worth investigating what types of challenges have been brought to entity linking because of language barriers. The top cross-lingual entity linking systems can be ranked at top 4 and 5 in the mono-lingual track, better than most mono-lingual entity linking systems.

REFERENCES

[1] James F. Allen. 1983. Maintaining Knowledge about Temporal Intervals. Communications of the ACM, November 1983, Volume 26, Number 11, pp. 832-843.
[2] Javier Artiles, Qi Li, Taylor Cassidy, Suzanne Tamang and Heng Ji. 2011. CUNY BLENDER TAC-KBP2011 Temporal Slot Filling System Description. Proc. Text Analysis Conference (TAC2011).
[3] Amit Bagga and Breck Baldwin. 1998 Algorithms for Scoring Coreference Chains. Proc. Resources and Evaluation Workshop on Linguistics Coreference.
[4] Chitta Baral, Gregory Gelfond, Michael Gelfond and Richard B. Scherl. 2005. Textual Inference by Combining Multiple Logic Programming Paradigms.
[5] L. Bentivogli, P. Clark, I. Dagan, H.T. Dang and D. Giampiccolo. 2010. The Sixth PASCAL Recognizing Textual Entailment Challenge. Proc. Text Analysis Conference (TAC2010) .
[6] Taylor Cassidy, Zheng Chen, Javier Artiles, Heng Ji, Hongbo Deng, Lev-Arie Ratinov, Jing Zheng, Jiawei Han and Dan Roth. 2010. CUNY-UIUC-SRI TAC-KBP2011 Entity Linking System Description. Proc. Text Analysis Conference (TAC2011) .
[7] Zheng Chen and Heng Ji. 2011. Collaborative Ranking: A Case Study on Entity Linking. Proc. EMNLP2011.
[8] Mark Dredze, Paul McNamee, Delip Rao, Adam Gerber and Tim Finin. 2010. Entity Disambiguation for Knowledge Base Population. Proc. NAACL2010.
[9] Noemie Elhadad, Regina Barzilay and Kathleen McKeown. 2002. Inferring Strategies for Sentence Ordering in Multidocument Summarization. JAIR, 17:35-55.
[10] Angela Fahrni and Michael Strube. 2011. HITS' Cross-lingual Entity Linking System at TAC2011: One Model for All Languages. Proc. TAC2011.
[11] William A. Gale, Kenneth W. Church and David Yarowsky. 1992. One Sense Per Discourse. Proc. DARPA Speech and Natural Language Workshop.
[12] Swapna Gottipati and Jing Jiang. 2011. Linking Entities to a Knowledge Base with Query Expansion. Proc. EMNLP2011.

[13] Prashant Gupta and Heng Ji. 2009. Predicting Unknown Time Arguments based on Cross-event Propagation. Proc. ACL-IJCNLP 2009.

[14] Xianpei Han and Le Sun. 2011. A Generative Entity-Mention Model for Linking Entities with Knowledge Base. Proc. ACL2011 Zellig Harris. 1954. Distributional Structure. Word , 10(23):146-162.

[15] Ludovic Jean-Louis, Romaric Resancon, Olivier Ferret and Wei Wang. 2011. Using a Weakly Supervised Approach and Lexical Patterns for the KBP Slot Filling Task. Proc. TAC2011.

[16] Heng Ji, Ralph Grishman, Hoa Trang Dang, Kira Griffitt and Joe Ellis. 2010. An Overview of the TAC2010 Knowledge Base Population Track. Proc. Text Analytics Conference (TAC2010).

[17] Heng Ji, Ralph Grishman, Dayne Freitag, Matthias Blume, John Wang, Shahram Khadivi, Richard Zens and Hermann Ney. 2009. Name Translation for Distillation. Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation.

[18] Zornitsa Kozareva and Sujith Ravi. 2011. Unsupervised Name Ambiguity Resolution Using A Generative Model. Proc. EMNLP2011 Workshop on Unsupervised Learning in NLP.

[19] Xuansong Li, Joe Ellis, Kira Griffitt, Stephanie M. Strassel, Robert Parker and onathan Wright 2011. Linguistic Resources for 2011 Knowledge Base Population Evaluation. Proc. TAC2011.

[20] Xiao Ling and Daniel S. Weld. 2010. Temporal Information Extraction. Proceedings of the Twenty Fifth National Conference on Artificial Intelligence.

[21] Paul McNamee, James Mayfield, Douglas W. Oard, Tan Xu, Ke Wu, Veselin Stoyanov and David Doermann. 2011. Cross-Language Entity Linking in Maryland during a Hurricane. Proc. TAC2011.

[22] Bonan Min and Ralph Grishman. 2012. Challenges in the TAC-KBP Slot Filling Task. Proc. 8th International Conf. on Language Resources and Evaluation. David Milne and Lan H. Witten. 2008. Learning to Link with Wikipedia. Proc. CIKM2008.

[23] Mike Mintz, Steven Bills, Rion Snow and Dan Jurafsky. 2009. Distant Supervision for Relation Extraction without Labeled Data. Proc. ACL-IJCNLP2009.

[24] Sean Monahan, John Lehmann, Timothy Nyberg, Jesse Plymale and Arnold Jung. 2011. Cross-Lingual Cross-Document Coreference with Entity Linking. Proc. TAC2011. Danuta Ploch. 2011. Exploring Entity Relations for Named Entity Disambiguation. Proc. ACL2011..

