When ACE met KBP[1]: End-to-End Evaluation of Knowledge Base Population with Component-level Annotation

Bonan Min[†], Marjorie Freedman[‡], Roger Bock[†], Ralph Weischedel[‡]
[†] Raytheon BBN Technologies, Cambridge, MA 02138
{bonan.min, roger.bock}@raytheon.com
[‡] USC/Information Sciences Institute, Marina del Rey, CA 90292
{mrf, weisched}@isi.edu

Abstract

Building a Knowledge Base from text corpora is useful for many applications such as question answering and web search. Since 2012, the Cold Start Knowledge Base Population (KBP) evaluation at the Text Analysis Conference (TAC) has attracted many participants. Despite the popularity, the Cold Start KBP evaluation has several problems including but not limited to the following two: first, each year's assessment dataset is a pooled set of query-answer pairs, primarily generated by participating systems. It is well known to participants that there is pooling bias: a system developed outside of the official evaluation period is not rewarded for finding novel answers, but rather is penalized for doing so. Second, the assessment dataset, constructed with lots of human effort, offers little help in training information extraction algorithms which are crucial ingredients for the end-to-end KBP task. To address these problems, we propose a new unbiased evaluation methodology that uses existing component-level annotation such as the Automatic Content Extraction (ACE) dataset, to evaluate Cold Start KBP. We also propose bootstrap resampling to provide statistical significance to the results reported. We will then present experimental results and analysis.

Keywords: Information Extraction, Knowledge Base Population, evaluation

1. Introduction

Automatically constructing a Knowledge Base (KB) of entities and relations from unstructured text, has long been a goal of Natural Language Processing (NLP). The task, named Knowledge Based Population (KBP), will unlock the huge potential in unstructured text for applications such as questions answering and web search.

Since 2012, the National Institute of Standards and Technology (NIST) has run the TAC Cold Start KBP evaluation, which measures performance of KBP. As the successor to the Message Understanding Conference (MUC) (Grishman and Sundheim, 1996) and Automatic Content Extraction (ACE) (Doddington et al., 2004) evaluations, Cold Start KBP evaluates a system's ability to automatically construct a KB from text. It uses a large corpus of 50,000-90,000 documents which have not gone through a careful selection process. In Cold Start KBP evaluation, a system is required to submit a KB [2] of entities and relations, constructed automatically from the corpus by algorithms.

How can one evaluate the quality of a KB? The Cold Start KBP evaluation[2] measures it by probing the KB with two types of queries: 1-hop (e.g., *which organization(s) is(are) founded by Bill Gates?)* or 2-hop (e.g., *in which city(-ies) is(are) the organization(s) founded by Bill Gates headquartered?*). The evaluation software traverses a KB and finds all answers to the 1-hop and 2-hop queries. Human annotators then annotate the correctness of a system answer by checking whether it is sufficiently justified in the source corpus. The process is performed over all submitted KBs[3].

While the Cold Start KBP evaluation directly measures end-to-end performance on KBP, it has several problems:

- The scores will vary by number of participants and the amount of answers they produced. Furthermore, the scores aren't comparable from year to year, therefore it is hard to measure progress.
- Given the high cost of the assessment process, the query set has typically been small relative to the schema size. For example, the 2016 query set contains only 317 queries - not a large number for 42 relation types.
- The evaluation suffers from severe pooling bias. Chaganty et al. (2017) show that the Cold Start KBP evaluation is significantly and systematically biased against systems that make novel predictions. For a system that does not participate in each year's evaluation, the pool is likely to not contain a significantly large fraction of correct answers. Therefore, recall will be significantly underestimated. Precision will also be estimated incorrectly because of novel answers that are not assessed.
- The assessment dataset is at the end-to-end (query-answer pair) level. It offers little for improving the components of a KBP system. A standard approach (Ji and Grishman, 2011) to KBP is to integrate a range of Information Extraction (IE) technologies including: named entity recognition, within document coreference, relation extraction, and cross document coreference. The KBP dataset cannot be used for (re)training any of the component level algorithms.
- The KBP assessment dataset annotation lacks component level annotation to support error analysis. A KBP system developer must trace the cause of an error. On the Slot Filling subtask alone, Min et al. (2012)

[1] Inspired by the movie "*When Harry Met Sally…"* in which two friends with drastically different personalities found each other to be the love of their life.

[2] The schema of the KB and the evaluation procedure, are defined in the task description, available at https://tac.nist.gov/2017/KBP/ColdStart/guidelines/TAC_KBP_2017_ColdStartTaskDescription_1.0.pdf

[3] A time-limited manual run is conducted and used to increase the size of the answer pool.

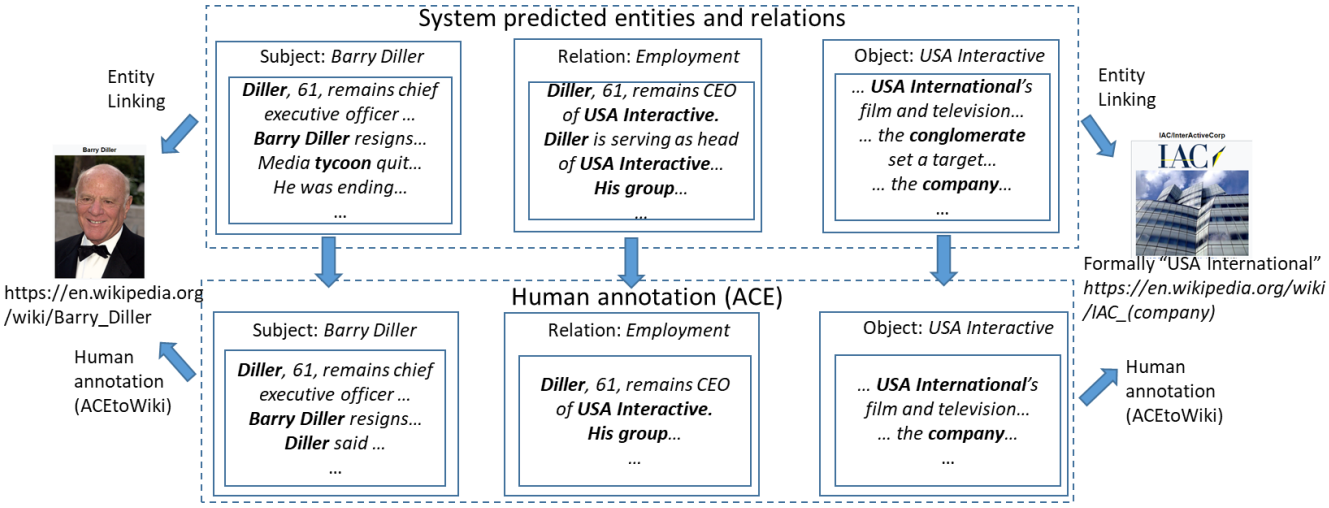


Figure 1 Aligning a system-predict relation to ground truth. The arrows show the direction of alignment.

shows that error analysis requires significant manual effort even with the help of KBP assessments.

- While it may seem that participants could accurately estimate the quality over the overall KB by measuring performance on traditional information extraction tasks (e.g. with standard relation extraction, coreference, and name-entity recognition datasets and metrics), experience has shown that improvements in an enabling technology do not translate to improvements in overall knowledge base quality. We hypothesize this is due to the differences in focus between sentence-by-sentence information extraction and a task that examines a corpus as whole. As an example, KB quality is impacted less by finding additional instances of the same very common fact, where as in sentence-by-sentence extraction finding an additional instance of a fact that has been seen in the previous sentence is weighted equivalently to finding a new fact.

To address these problems, we propose a novel evaluation method that uses existing information extraction resources such as the ACE training corpus (Walker et al, 2016) to evaluate Cold Start KBP. It has no pooling bias, does not rely on carefully selected queries, and can be used to measure progress since no additional annotation is required for a new system. The dataset has component-level annotation, therefore it supports fine-grained analysis of errors and can be used for training/improving IE components. Furthermore, we augment the method with bootstrap resampling to provide statistical significance. We present experiments and analysis.

2. Related Work

The TAC Cold Start KBP evaluation provides a corpus of 50,000 to 90,000 documents. A system is expected to produce a KB of 5 entity types and 42 relation types defined in the TAC schema[2]. The evaluation software probes each submission KB with 1-hop or 2-hop queries and obtains a set of answers. A separate time-limited human answer-finding round is also conducted to add more answers. The pooled answer set, accompanied with justification text in the original corpus, are provided to human annotators to be assessed as correct, incorrect, or redundant. A system is measured by the precision and recall of its answers to queries, using the pool of assessed answers. The evaluation is subject to pooling bias. Lacking identification of which component the error stems from, the annotation is not very useful for analyzing the error nor can it be used to improve or (re)train new component-level models for system improvement.

Recently, Chaganty et al. (2017) proposed an on-demand evaluation framework for Cold Start KBP. Observing that newly developed systems suffer from significant pooling bias, they proposed to use crowdsourcing to annotate newly found answers on-demand, and an importance sampling strategy for unbiased evaluation. Although the collected annotation contains EDL (entity discovery and linking) and Slot filling annotation, it still is not sufficiently useful for fine-grained error analysis such as mention tagging, coreference, etc. Moreover, the annotation is collected at the end-to-end level; therefore it is not straightforward to use it to train component-level algorithms. Furthermore, the cost for each new system is about $300. The total cost could potentially be very large if many systems need to be evaluated on-demand, e.g. to support variations in parameters and/or algorithms.

3. A Novel Evaluation Method

Resources We use the ACE 2005 English training corpus (Walker et al., 2006) as the evaluation corpus. The ACE corpus consists of articles from weblogs, broadcast news, newsgroups, and broadcast conversations; it is annotated exhaustively with mentions, coreference, and relations. To augment the ACE corpus with entity linking annotation, we use ACEtoWiki[4] (Bentivogli et al., 2010), an auxiliary dataset which extends the English ACE 2005 corpus annotation with ground-truth links to Wikipedia.

The ACE dataset is widely used to evaluate IE components such as named entity recognition, within-document coreference, and relation extraction. Since there is a plethora of work (Nadeau and Sekine, 2007; Luo, 2005) on component-level evaluations, we will refer interested readers to these papers. As a benchmark dataset for IE components, ACE helps KBP system developers to find places to improve and understand

[4] We manually remove links to under-specified pages or links to groups of entities, e.g., links to *People*, *Presidents of the United States*, *Country*, *Politician*, *etc*. This resulted in removal of 14% links.

where the errors are. The ACE dataset has also been used extensively for training entity (Nadeau and Sekine, 2007) and relation tagging models (Zhou et al., 2005).

We will focus on the end-to-end evaluation and describe how to use the ACE and ACEtoWiki augmentation to evaluate KBP. At a high level, the idea is to first align system predicted relation triples *{<subject, relation, object>}* to the ground truth triples {*<subject', relation', object'>}*, and then generate relation paths of a single or multiple hops with each hop being a relation triple. The resulting ground truth relation paths, can be used as reference for measuring how accurate a system is (precision) and its coverage (recall). Figure 1 shows an example of the process to align system predicted relation triples to the ground truth. We need to perform alignment at the following levels:

Document-level entity: An alignment is found if *>P%* (*P=50* in the experiments) named mentions in a system-predicted entity cluster can be aligned to one of the mentions in a corresponding ground truth entity. For example, a system-generated *Barry Diller* entity will be aligned to a ground truth entity *e* if more than 50% of the system-tagged mentions are in *e*. As illustrated in Figure 1, we perform entity alignment for the relation's subject *Barry Diller*, and object *USA Interactive*.

Entity Linking: For each entity, we use ACEtoWiki dataset to find a Wikipedia page for each named mention (if the page exists). The Wikipedia page of the most named mentions will be used as the ground truth page for the entity. For the names that are unlinkable to any Wikipedia pages, we cluster them by exact string match. This generates a unique corpus-level entity for one or multiple coreferential document-level entities.

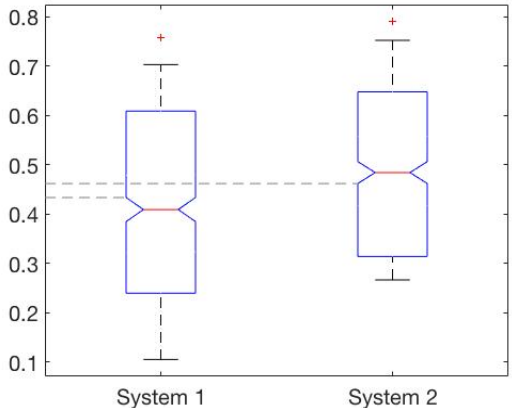


Figure 2 Illustration of scores intervals and comparing of system performance. The box shows 95% (top) and 5% (bottom). Outliers (red dots) are above the top horizontal line or below the bottom horizontal line. Medians are represented with red lines.

Relation: A system-predicted relation can be aligned to a ground truth relation *r*, if its subject, relation type and object all can be aligned to the corresponding fields of *r*. We found a relation type match if the system-predicted type is exactly the same type in the ground truth. We didn't enforce relation provenances being equal since we focus on corpus-level end-to-end KBP.

Relational paths: To support evaluation of 1-hop, 2-hop, and multi-hop queries (e.g., answering a 2-hop question *in which city(-ies) the organization(s) founded by Bill Gates is(are) headquartered?*), we generate *n*-hop relational paths in the form of $< e_1, r_1, r_2, \dots, r_n, e_{n+1} >$. e_1, e_{n+1} are head and tail entities. $r_1, r_2 \dots$ are relation types. We define one-hop paths (relation triples) as R_1:

$$R = R_1 = \{< e_1, r_1, e_2 >\}$$

Further, we define two-hop paths R_2 as the following:

$$R_2 = R_1 \hat{\times} R = \{< e_1, r_1, r_2, e_3 > | \forall < e_1, r_1, e_2 >, < e_2, r_2, e_3 > \in R\}$$

and n-hop paths R_n:

$$R_n = R_{n-1} \hat{\times} R = \{< e_1, r_1, r_2, \dots, r_n, e_{n+1} > | \forall < e_1, r_1, e_2, r_2, e_3, \dots, r_{n-1}, e_n > \in R_{n-1}, \forall < e_n, r_n, e_{n+1} > \in R\}$$

We generate two sets of multi-hop paths with the above mentioned equation: R_n from ground truth annotation, and R'_n from the system predicted tuples. We compare R'_n to R_n, and calcualte preicsion, recall and F1:

$$P = \frac{|R_n \cap R'_n|}{R'_n}, R = \frac{|R_n \cap R'_n|}{R_n}, F1 = \frac{2PR}{P+R}$$

Bootstrap re-sampling: The ACE English dataset contains 596 documents. Similar datasets (e.g., Rich ERE (Song et al., 2015)) with entity and relation annotation contains a similar number of documents. To show the statistical variance of the scores as well as to show whether any increase in scores is statistically significant, we propose bootstrap resampling (Efron and Tibshirani, 1994), which has been applied in tasks such as measuring Machine Translation performance (Koehn, 2004).

Bootstrap resampling works as follows: Assume we can only measure performance with *n* (*n=596* for ACE) documents, randomly drawn from some large corpus in the ideal world (on which it would be too expensive to annotate exhaustively). We compute precision, recall and F1 with the above-mentioned method on the *n* documents. We could sample another test set of *n* documents from the original *n* documents with replacement, and compute the scores again. We repeat this for a sufficiently large number of times (e.g., 1000 times) and produce many measures of the performance. These sampled scores can be used for

- Estimating an interval [*a*, *b*] which approaches the 90% confidence interval for scores of test set of size *n*. To do so, we will sort the scores and then drop the top 5% and bottom 5%. We show the interval as a box as illustrated for *System 1* in Figure 2. Outliers are above the top horizontal line or below the bottom horizontal line.
- Estimating whether an improvement in score is statistically significant. To do so, we construct a pair of bootstrap resampling scores, one for a baseline system *System 1*, and the other for an improved system *System 2*. Both are illustrated in Figure 2. The notch of a box shows the confidence interval which is normally based on the median$\pm 1.15 * IQR/\sqrt{n}$ (the interquartile range (IQR) is the 25 to 75 percentage). Notches (Chambers et al., 1983) are useful in offering a rough guide to significance of difference of medians; if the notches of two boxes do not overlap, this offers evidence of a statistically significant difference between (95% confidence) the medians. The two notches in Figure 2

don't overlap, it shows the improvements (as defined by the improvement in median) is statistically significant.

4. Implementation and Experiments

As described in Section 3, we use the ACE English dataset as the evaluation dataset. We use relation subtypes since subtypes represent concrete relations such as *Employement*, *Subsidiary* instead of their categorical counterparts (types defined in ACE) such as *Organization-Affiliation* and *Part-Whole*. We do not include *User-Owner-Inventor-Manufacturer*, *Citizen-Resident-Religion-Ethnicity*, and *Lasting-Personal* since the meaning of these relations are not clear - each can be further divided into finer-grained types.

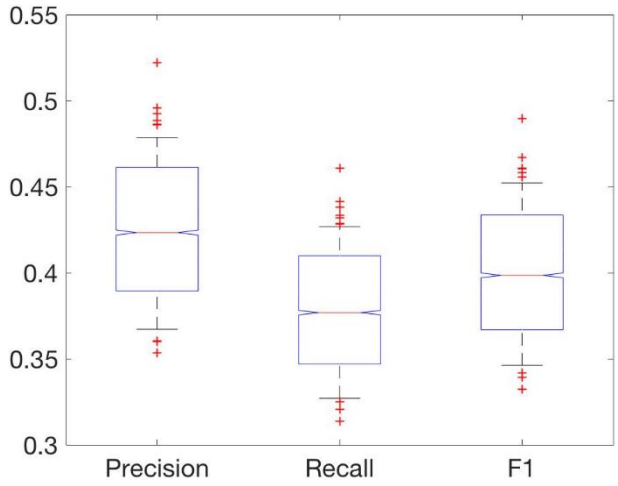
Table 1 System performance on the ACE corpus.

	Precision	Recall	F1
1-hop	0.427	0.38	0.402
2-hop	0.24	0.156	0.188
3-hop	0.103	0.09	0.095

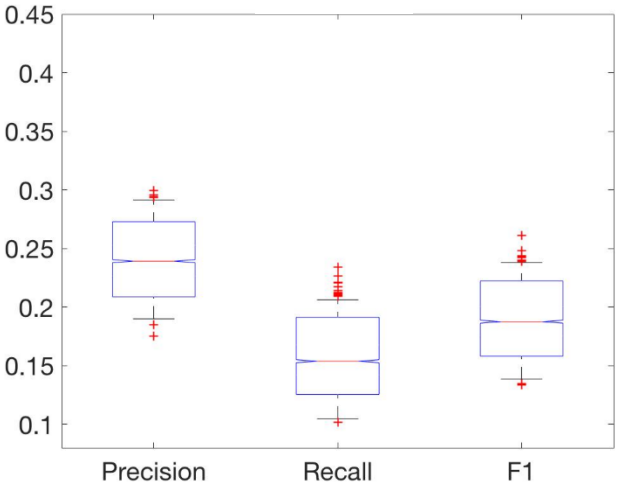
We apply a state-of-the-art Cold Start KBP system (Min et al., 2017; Min and Freedman, 2016) on the ACE English document sets. Since TAC KBP has different type sets for relations, we only apply the ACE relation extraction system in Min and Freeman (2016)[5] to support the ACE relation schema.

Table 1 shows the performance on the set of 596 ACE documents without any sampling approach. We measure performance up to three hops (e.g., "*Where is the organization Bill Gates's mother works for?*"). Both precision and recall drop as we add hops. This shows the known problem of error multiplication in Cold Start KBP: Errors accumulate along the paths and renders the end results less precise and have lower coverage.

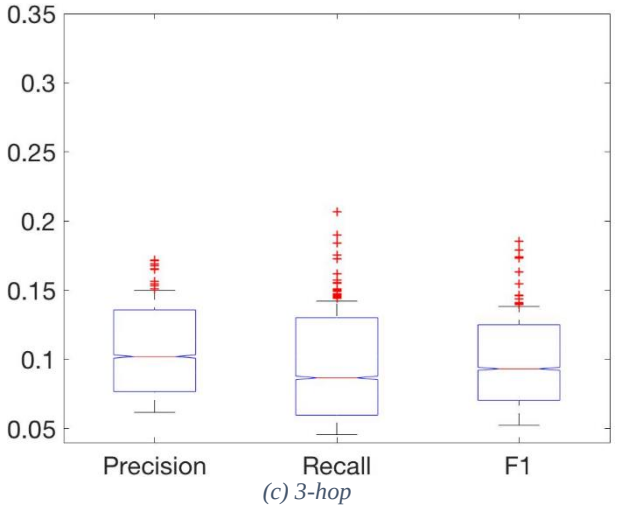
We further experimented with bootstrap resampling. We ran the experiment 1000 times, sampling documents with replacement. Figure 3 shows the results on 1-hop (Figure 3a), 2-hop (Figure 3b), and 3-hop (Figure 3c) respectively. Similar to Table 1, precision, recall and F1 also decrease as hop increases. The variance (length of boxes as well as bars indicating outliers) is not very large. This indicates that we could obtain very accurate estimates for each measure with a large sampling size (1000 for our experiments). In addition, the notches are very small (<1 point). This offers an accurate way to measure statistically significant improvements to the system – for scores obtained with bootstrap resampling for a new system, if the notch of the new scores didnot overlap with the current notch, it indicates a statically significant difference. The small notches on all experiments show that we could measure statistically significant improvements fairly accurately.



(a) 1-hop



(b) 2-hop



(c) 3-hop

Figure 3 System performance with bootstrap resampling on 1-hop (top), 2-hop (middle), and 3-hop (bottom). The scores are generated with 1000 bootstrap resampling runs. The top and bottom lines of the boxes show 95% and 5% percentile respectively. The notch (though very small due to the low variance in median scores) shows confidence intervals of median scores.

5. Conclusion and Future Work

We present a novel method for evaluating end-to-end Knowledge Base Population with component-level annotation. Our method makes use of existing component-

[5] One of the main relation extraction components in Min and Freedman (2016) is a set of statistical models trained with the ACE training dataset. Its decoding output can be mapped into KBP relation types, given the similarity between ACE and KBP. We use the unmodified ACE type output to support direct assessment on the ACE dataset.

level annotation such as ACE. It also includes bootstrap resampling approaches for measuring statistical significance of the results. Our next step is to apply the approach to other datasets such as the rich ERE (Entity, Relation and Events) (Song et al., 2015) annotation dataset.

6. Acknowledgements

This work was supported by DARPA/I2O Contract No. FA8750-13-C-0008 under the DEFT program. The views, opinions, and/or findings expressed are those of the author(s) and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government. This document does not contain technology or technical data controlled under either the U.S. International Traffic in Arms Regulations or the U.S. Export Administration Regulations.

7. Bibliographical References

Luisa Bentivogli, Pamela Forner, Claudio Giuliano, Alessandro Marchetti, Emanuele Pianta, Kateryna Tymoshenko. Extending English ACE 2005 Corpus Annotation with Ground-truth Links to Wikipedia. In Proceedings of COLING 2010 Workshop on "The People's Web Meets NLP: Collaboratively Constructed Semantic Resources", Beijing, China, August 28, 2010.

Arun Chaganty, Ashwin Paranjape, Percy Liang and Christopher D. Manning. Importance sampling for unbiased on-demand evaluation of knowledge base population. In Proceedings of EMNLP 2017.

John M. Chambers, William S. Cleveland, Beat Kleiner, and Paul A. Tukey. 1983. "Comparing Data Distributions." In Graphical Methods for Data Analysis, 62. Belmont, California: Wadsworth International Group.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, Ralph Weischedel. The Automatic Content Extraction (ACE) Program-Tasks, Data, and Evaluation. In Proceedings of LREC 2004.

Bradley Efron and R. J. Tibshirani. 1994. An Introduction to the Bootstrap. CRC Press.

Ralph Grishman and Beth Sundheim, Message Understanding Conference - 6: A Brief History. In Proceedings of the 16th International Conference on Computational Linguistics, Copenhagen, Denmark, 1996.

Paul McNamee, Hoa T. Dang, Heather Simpson, Patrick Schone, Stephanie Strassel. An Evaluation of Technologies for Knowledge Base Population. In Proceedings of LREC 2010.

Bonan Min and Marjorie Freedman. BBN's 2016 System for Cold Start Knowledge Base Population. In Proceedings of the Text Analysis Conference (TAC).

Bonan Min, Marjorie Freedman and Talya Meltzer. Probabilistic Inference for Cold Start Knowledge Base Population with Prior World Knowledge. In Proceedings of EACL 2017.

Bonan Min and Ralph Grishman. Challenges in the Knowledge Base Population Slot Filling Task. In Proceedings of LREC 2012.

Heng Ji and Ralph Grishman. Knowledge base population: Successful approaches and challenges. In Proceedings of ACL 2011.

Philipp Koehn. Statistical Significance Tests for Machine Translation Evaluation. In Proceedings of EMNLP 2004.

David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. Lingvisticæ Investigationes, Volume 30, Issue 1, 2007, pages: 3 – 26.

Xiaoqiang Luo. On coreference resolution performance metrics. In Proceedings of HLT-EMNLP 2005

GuoDong Zhou, Jian Su, Jie Zhang and Min Zhang. Exploring various knowledge in relation extraction. In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics

8. Language Resource References

Zhiyi Song, Ann Bies, Stephanie Strassel, Tom Riese, Justin Mott, Joe Ellis, Jonathan Wright, Seth Kulick, Neville Ryant and Xiaoyi Ma. From Light to Rich ERE: Annotation of Entities, Relations, and Events. In Proceedings of the 3rd Workshop on EVENTS at the NAACL-HLT 2015.

C. Walker, S. Strassel, J. Medero, K. Maeda. ACE 2005 multilingual training corpus. Linguistic Data Consortium, Philadelphia, 2006.