CEUR-WS.org/Vol-2127/paper3-kg4ir.pdf

Summarizing Entities using Distantly Supervised Information Extractors

Travis Wolfe[1,2], Annabelle Carrell[1], Mark Dredze[1], and Benjamin Van Durme[1]

[1]Johns Hopkins University, Baltimore MD, USA
[2]Google, Mountain View, CA, USA*

Abstract

We describe entity summarization, the task of producing informative text summaries for an entity described across multiple documents in a collection. Existing (multi-)document summarization techniques are applied as baselines to this task, which we generalize to allow for joint information extraction and summarization. Through user evaluations across a variety of approaches, we discover what is most preferred in an entity summary. In particular we find that distantly supervised information extractors lead to significant improvements over lexical approaches, demonstrating the utility of extraction technologies for a task other than knowledge base population.

1 Introduction

Many professional knowledge workers (reporters, financial analysts, scientists, politicians, etc), consume large volumes of written reports every day. Their information needs often involve questions about a given entity. Single and multi-document summarization techniques compress texts which discuss many entities, not all of which are equally important to a reader with an entity-centric information need. Automatic methods for assisting knowledge workers in doing directed exploration of large amounts of textual information are lacking.

Topic models are a popular method for corpus exploration, aiming to help find subsets of a document collection most salient to a user's needs. However, topics offer little in terms of structure which can be used to navigate and search through documents for specific entities. Knowledge Base Population (KBP) [19, 15] aims to automatically discover the entities and useful relations mentioned in text, to be organized into a knowledge base (KB) supporting structured queries. However, while KBP methods continue to improve, they remain imperfect, requiring the final information presented to a user to include provenance information, typically the text that a fact was extracted from. Systems that behave this way share much in common with text summarization methods, but with more explicit models of information (entities, facts, relations).

The central question in entity summarization is about the value of information: what should go into a summary about an entity? Most previous work in document summarization has addressed this question at the lexical level. Here we propose two new units of information about entities: the presence of *related entities* and of *facts about an entity*. We show that: (1) lexical methods are correlated with methods which track the presence of related

*Work was completed while at Johns Hopkins University.

Copyright © by the paper's authors. Copying permitted for private and academic purposes.

In: Joint Proceedings of the First International Workshop on Professional Search (ProfS2018); the Second Workshop on Knowledge Graphs and Semantics for Text Retrieval, Analysis, and Understanding (KG4IR); and the International Workshop on Data Search (DATA:SEARCHâĂŹ18). Co-located with SIGIR 2018, Ann Arbor, Michigan, USA – 12 July 2018, published at http://ceur-ws.org

entities, but are subject to ambiguities often ignored in document summarization, and (2) modeling facts about entities produces better summaries than lexical models of information content.

Wikipedia is a popular entity-centric knowledge resource, but it only covers public information and has lower coverage on less popular entities. Methods which extract facts and summaries from arbitrary natural language sources (Wikipedia being just one) have a natural advantage in coverage. We use *Wikipedia infoboxes* as training data for two aspects of our method with the goal of generalizing over entities mentioned in arbitrary natural language. First, we use the types of facts[1] in infoboxes as a signal as to what facts may people find important to quickly glean about an entity. Second, we use the tokens of facts in infoboxes as distant supervision [2, 21] for fact extractors which can nominate information rich sentences to include in a summary. We evaluate the generated summaries via human preference judgments, finding that they are significantly more informative than lexical methods drawn from other summarization tasks.

2 Entity Summarization Task

The goal of summarization in general is to take a collection of documents such as news articles, books, financial reports, scientific papers, or encyclopedia entries and produce a short and informative summary. Entity summarization is the case where the summary has an entity which serves as its subject.

The first step in entity summarization is finding a suitable set of source texts related to the subject. There are a variety of methods for finding source texts depending on how the subject is specified, such as providing an article about the subject [16], or free text queries [26]. In this work we address the basic case where a subject is specified as a disambiguated entity identifier and rely on prior work to identify all articles in a news corpus which mention the subject, c.f. §6.

Given a set of texts which contain the subject has been nominated, the text must be analyzed at the mention level to identify places where the subject is mentioned, which is related to work on entity linking [19], cross document coreference resolution [27, 23], and entity mention search [7]. Sentences which contain mentions of the subject are the basic inputs needed for an extractive summarization techniques, which we use as a baseline in §6. Richer use of mentions of the subject, such fact extraction §4 or related entity analysis §5 can be used to extract deeper features used to quantify sentences' information utility.

Finally, a summarization model must aggregate all of the subject mentions into a summary with length constraints. This step involves weighing a variety of concerns for what makes a good summary. Sentences in the summary must be relevant to the subject. This distinguishes entity summarization from (multi-) document [10] summarization where all the input sentences are assumed to be relevant to the output. Similar to other summarization techniques [22], the summary should not be repetitive, or alternatively it should fit as much information as the textual limit allows [4, 9]. The summary should be coherent and fluent, for example ensuring that all pronouns have easily interprettable referents. We discuss our summarization model in §3.

3 Summarization Model

Our approach begins by assuming a baseline notion of a sentence's relevance to the subject: whether the subject is explicitly mentioned in the sentence. Given this (potentially large) subset of sentences, our model is an extractive summarization model which focuses on methods for quantifying the value of information contained in sentences and the uncertainty associated with extracting it. For now we will discuss the model abstractly in terms of *concepts*, or discrete pieces of information mentioned in the source sentences. Later in §4 and §5 we will describe implementations for these concepts.

The model described in [9] provides an extensible framework for extractive summarization based on the appearance of concepts, each of which has a *utility* to the reader and a list of locations it was mentioned. Their model is declarative: it describes an objective to be maximized rather than an heuristic algorithm for finding good solutions [18]. Modern solvers can perform efficient inference, making this model very scalable.

We extend the model of [9] in order to allow for uncertainty over the presence of concepts. Lexical models of information content have the advantage of lacking this uncertainty, but they cannot quantify information content in terms of facts which which are inferred through natural language understanding. To handle this uncertainty our model incorporates MAP inference over classifier predictions.

Both models find a summary of maximal utility which respects a length constraint using similar integer linear program (ILP) formulations presented in Figure 1, where:

[1]We invoke the type/token distinction for "facts" where a fact *type* is a relation like BornIn(x: Person, y: Location), and a fact *token* is a particular fact like BornIn(Justin_Trudeau, Ottawa_Canada).

$$\max_{s,c} \sum_i w_i c_i \tag{1}$$
$$s.t. \sum_j l_j s_j \leq L \tag{2}$$
$$s_j Occ_{ij} \leq c_i \tag{3}$$
$$\sum_j s_j Occ_{ij} \geq c_i \tag{4}$$
$$c_i \in \{0,1\} \ \forall i \tag{5}$$
$$s_j \in \{0,1\} \ \forall j \tag{6}$$

$$\max_{s,c,e} \sum_i w_i c_i + \alpha e_{ij} \log p_{ij} \tag{7}$$
$$s.t. \sum_j l_j s_j \leq L \tag{8}$$
$$s_j e_{ij} \leq c_i \tag{9}$$
$$\sum_j s_j e_{ij} \geq c_i \tag{10}$$
$$c_i \in \{0,1\} \ \forall i \tag{11}$$
$$s_j \in \{0,1\} \ \forall j \tag{12}$$
$$e_{ij} \in \{0,1\} \ \forall i,j \tag{13}$$

Figure 1: Left: the model of [9], Right: an extension which includes uncertainty over concept occurrence/extraction (this work).

- w_i is a weight, or utility, for including the i^{th} concept in the summary
- c_i is a binary variable indicating whether the i^{th} concept is included in the summary
- Occ_{ij} is a binary *value* indicating that sentence j contains concept i
- e_{ij} is a binary *variable* indicating that sentence j contains concept i
- p_{ij} the extraction models probability estimate that sentence j contains concept i
- s_j is a binary variable indicating that the summary contains sentence j

We use [9] as a baseline with word bigrams as concepts. w_i is computed in the same fashion as computing a tf-idf vector for bigrams. [9] report that they didn't need to use idf weighting, but in this work it appeared to help during development. One reason could be that we are summarizing a great deal more material than in the TAC summarization task. There, systems received 10 articles of source text, whereas here we take all mentions of an entity in ClueWeb09 (which can be over 1 million tokens for common entities). With so many choices, optimal solutions often include odd sentences with abnormal amounts of frequent bigrams which do not contribute to good summaries.

In the next two sections we will describe two new concept definitions. We found that that using more than one type of concept in the same model made it difficult to balance the influence of each concept type using linear coefficients per concept type. There is significant variation across summary subjects of what the utility scale is for one type of concepts versus another. We found that we could make our model attendant to each type of concept for most queries by normalizing each concept to have unit weight. More precisely, if a concept type is a set of indices T, we ensure that $\sum_{i \in T} w_i = 1 \ \forall T$.

Additionally we found that in some cases the costs outweighed the utility of the concepts when using our default of $\alpha = 1$. In other cases this might be worth knowing (i.e. "we can't infer much about this entity given their mentions"), but in our case we would like the best summary which hits the length limit. If our system produces an empty summary, we halve α and re-run optimization.

4 Infobox Distant Supervision

A natural way to model the informativeness of an entity summary is to count the number of facts contained in the summary. Some facts matter more than others, e.g. where a person was born is more important than whether a university has an even or odd number of students. To determine which facts are informative we use *Wikipedia infoboxes*, a curated set of facts about an entity. These facts cover a wide range of types of entities and express things like which countries were affected by an earthquake, what company manufactures a drug, or where a person was born. Because they consume two finite resources, the time required by a person to write them down and the space at the top of a Wikipedia article, it is safe to conclude that these facts have utility.

The difficulty in using facts as concepts in our summarization model is that our systems can't observe facts directly, they must be extracted from text, requiring relation extractors for infobox facts. Therefore, we perform textual relation extraction to discover concepts that are worth including in our summaries. We build relation extractors using distant supervision. As relation extraction relates to the utility and costs described in §3, we assume each mention of a relation (infobox fact) has unit utility and the extraction cost is the extractor's surprisal (computed using the MAP estimate of precision).

We train relation extractors using distant supervision using Wikipedia infoboxes from DBpedia as a source of true facts (our KB). One goal of this work is to ask whether relation extractions are a useful signal in an entity summarization system. As such, we implement a simple and efficient distant supervision method which is similar to but distinct from a few previous works.

When training a relation extractor via distant supervision, instance (sentence) level annotations are not available. Successful distant supervision methods are ones which do a good job inferring which sentences *may* contain a relation mention versus which *actually do*. Some approaches address this with generative latent variable models [28]. Others make plausible but conservative assumptions about the data such as "at least one aligned sentence commits to a fact" [14]. Another category of methods use feature pooling (over sentences) to get around the sentence-level inference task. Feature pooling can take the form of unioning binary features on sentences into binary features on entity pairs [24] or neural network approaches like taking the pointwise mean or max of learned features [30]. There has been extensive work on distant supervision for relation extraction [14, 28, 24, 30] inter alia. In this work we extend the unioned binary features aspect of the model of [24], but couple it with a feature induction technique to find more linguistically plausible and expressive extractors based on feature conjunctions. Due to the conjunctive and rich nature of these features, we distinguish them by calling them *extractors*.

Extractors are pairs of a relation and a syntactic fragment, similar to the shortest dependency fragments used in [24]. Generalizing shortest paths to syntactic fragments (sets of dependency edges) allows extractors to detect the presence of linguistic phenomenon like negation and quantification which make extractions problematic. Because there are many possible extractors and enumerating and scoring all of them is computationally prohibitive, we use a feature induction technique to grow full extractors from shortest dependency paths most frequent in positive distantly supervised examples. This induction technique approximately finds the best extractors as measured by pointwise mutual information between extracted entity pairs and facts for an extractor's relation.

We combine the predictions of extractors using a precision-ranked-rules framework which has been successfully employed on tasks like coreference resolution named entity recognition, and part of speech tagging. We estimate each extractor's confidence/surprisal using a beta-binomial model where the observations are the same as for computing PMI, co-occurrences of extracted entity pairs and facts in the KB. At prediction time, our model assigns the highest probability of a relation which was supported by *any* extractor which fires with that relation.

5 Related Entities as Concepts

We introduce a new concept definition: *related entities*. We started with lexical concepts, which are trivial to extract and have many ambiguity and sparsity issues, and then moved to factual concepts, which are difficult to extract but closely track information content. The purpose of related entities is to be a compromise. Entity disambiguation methods are more reliable than relation extraction methods and offer some of the information that factual concepts capture.

Take the example of summarizing a company. Their most related entities will be companies they do business with, high-level employees, products they make, and relevant regulatory bodies. Even if the summarization model does not know what the relation between the query and these related entities is, it can still assign them high utility on account of their association. Related entities are also useful for disambiguation. For example, if there were two "Michael Jordan" entities, a powerful disambiguation method (for summarization) is to include mentions of "Scottie Pippen" or "David Blei".

This leads to the question of how to infer what entities are related to a query from text. In cases where the text is a part of communications like emails which have a sender and receivers, a graph of associated entities is fairly easy to create. We are primarily interested in sources which do not have senders and receivers, but we have found that the presence of two entities in the same sentences is a good source of evidence of association. Frequency alone can be misleading though. For example news agencies often report on many entities, and are

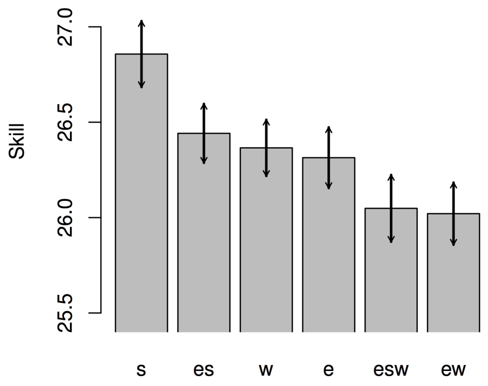

Figure 2: System quality as inferred by TrueSkill.

h_0 *vs* h_1	$h_0$wins	$h_1$wins	$p(x \mid \mu_0 = \mu_1)$
W *vs* V	135	116	0.256
EW *vs* EV	112	83	0.00780
ESW *vs* ESV	134	93	0.00779
*W *vs* *V	381	292	0.000682

Table 1: Comparisons between systems which use all words (W) vs words which don't overlap with a named entity (V).

thus mentioned frequently but are not associated with the entities they report about. We found that the same $idf(i)$ correction applied to lexical concepts works well for related entity concepts too.

Given the utility of related entity concepts, we set the extraction cost for each mention of a related entity equal to the number of words between the query and related entity mentions divided by 10.

6 Experiments

We hypothesize that using infobox relations as concepts will produce the most informative summaries out of the concept types outlined. Our experiments compare summaries containing infobox relations with summaries containing other concept combinations, defined as different systems.

A system is a set of concept definitions. We denote the baseline method of word bigrams as W, the infobox relation method as S (§4), and the related entities method as E (§5). Systems with two types of concepts are denoted with concatenation, e.g. S + E = SE. In total there are 7 systems: [E, S, W, ES, SW, EW, ESW].

We constructed a dataset of entities taken from the ClueWeb09 dataset [3] with the entity linking annotations provided by FACC1 [8]. This data has over 340m documents and 5.1b entity mentions which are resolved to Freebase MIDs. We partition our train, dev, and test sets by entity. We partition all entities in FACC1 into buckets by log frequency in order to include a balanced set of rare and frequent entities. For each sufficiently-sized bucket we selected 3800 train, 100 dev, and 100 test entities. We generate summaries of length $L \in \{20, 40, 80, 160\}$ words for each entity and average our results over all performance at every length.

To evaluate our methods we adopt methods from machine translation evaluation [25] which are based on human comparisons of system outputs. For each entity in our evaluation set, each summary length, and each pair of systems which produce a summary, we present Amazon Mechanical Turk annotators with a choice of which summary is more informative.

Given pairwise system preferences, we arrive at statistical inferences in two ways. If there are many systems competing, we use the TrueSkill algorithm [13] to infer the *skill*, a parameter which determines the likelihood that any one system will produce a more informative summary than another system, which we use to rank systems. If there are only two systems, we prefer the χ^2 test for better statistical efficiency.

7 Results

Figure 2 shows the skill inferred for every system. Defining concepts as infobox facts produces the most informative summaries. This is the only statistically significant result using W as the baseline.

Using related entities (E) produces summaries with comparable quality to word ngrams (W). Consistent with properties we saw during development: when using W concepts, the highest utility concepts, i.e. ngrams with high tf-idf weight, tend to be named entities. This indicates that most of the information that the E and W models are trying to preserve is the same. This may explain the fact that models which have both E and W components perform the worst: in these cases the model is redundantly rewarded for including the same information, thus not spending its word budget in the most informative way.

We can decompose the set of concepts which W is trying to preserve into entity concepts and non-entity concepts. While E is a natural replacement for the entity concepts in W, it is unclear how useful the non-entity concepts are. To quantify this, we defined a new set of concepts V which are the ngram concepts which do not overlap with any entity mention. We treat them the same as W concepts and weigh them via tf-idf. This information should be orthogonal to entities, indicating that perhaps EV might have fewer duplication problems compared to EW.

In Table 1 we performed a head-to-head evaluation between systems which use W concepts and the corresponding system with V instead. We produced a summary for both concept definitions and asked annotators which they preferred when the summaries were different. We see that in every case the V model performs worse, most frequently by statistically significant amount. This is further evidence that ngrams are not a sufficient method for modeling information beyond what named entities already capture. This highlights the effectiveness of the S concept definitions: S uses only non-entity information and performs significantly better than the strong entity baseline.

8 Related Work

There is a related line of work on entity summarization where the output is a set of facts or triples in a knowledge base [5, 12, 29]. These methods compute the importance of facts connected to a knowledge base entity using graph-based methods. This work focuses on producing textual summaries rather than triples and uses repetition in text or presence in an infobox as an orthogonal signal for a concept or fact's utility. Additionally this work is concerned with extracting facts from text rather than ranking facts in a KB. The TREC Entity track (2009-2011) [1] addressed the task of finding related entities to a query entity, which was treated as its own information retrieval task. Our work shows that related entities are an effective means of producing summaries. [11] performed TAC topic-based summarization using links into Wikipedia to determine how popular or salient an entity is, and thus how much utility should be assigned to including it in a summary. This method of estimating utility is directly applicable to our summarization model and is similar to our definition of related entities, but we do not assume that one has access to Wikipedia logs to determine popularity. [20] created entity-centric summaries for companies on Twitter, where the goal was highlight positive and negative sentiment, using hashtags as concepts. Similarly, [17] performs sentiment analysis and summarization jointly to create "micro reviews" on services like Yelp. [6] construct entity graphs similar to this work, but their goal is to summarize a path of entities in this graph rather than particular nodes. That work does not explicitly model concepts or produce summaries of a given length and only included a small scale evaluation of 27 entity chains.

9 Conclusion

We have investigated how existing extractive summarization methods can be adapted to produce *entity summaries* and compared the effectiveness of these methods to novel entity-centric concept definitions. To enable summarization in the presence of noisy extractors, we described a new ILP model which jointly performs MAP inference over extractions and summarization choices. When combined with our concept definitions this significantly outperformed document-centric extractive summarization baselines, as measured through a evaluation techniques from machine translation which make evaluating summarization techniques cheaper and more statistically powerful.

Distant supervision of Wikipedia infobox facts allowed us to train extractors which found useful information to incorporate into summaries, outperforming all other methods. Our experiments indicate that entity co-occurrences explain most of the signal that previous methods had exploited. By modeling entities explicitly through linking, we better weigh the importance of entities and avoid concept duplication and ambiguity issues introduced when only using words. For non-entity signal, we showed that our distant supervision model greatly outperforms previous methods. Together these results illustrate that while the community has not perfected automatic knowledge base construction (converting documents into fully structured repositories), existing technologies can be used today in helping users isolate text content that they will find worth reading.

Acknowledgement

This work was supported in part by the Human Language Technology Center of Excellence (HLTCOE) and the DARPA LORELEI program.

References

[1] Balog, K., Serdyukov, P., Vries, A.P.d.: Overview of the trec 2010 entity track. Technical report, DTIC Document (2010)

[2] Bunescu, R., Mooney, R.: Learning to extract relations from the web using minimal supervision. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Prague, Czech Republic, Association for Computational Linguistics (June 2007) 576–583

[3] Callan, J., Hoy, M., Yoo, C., Zhao, L.: Clueweb09 data set (2009)

[4] Carbonell, J., Goldstein, J.: The use of mmr, diversity-based reranking for reordering documents and producing summaries. In: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, ACM (1998) 335–336

[5] Cheng, G., Tran, T., Qu, Y. In: RELIN: Relatedness and Informativeness-Based Centrality for Entity Summarization. Springer Berlin Heidelberg, Berlin, Heidelberg (2011) 114–129

[6] Chhabra, S.: Entity-centric summarization: Generating text summaries for graph snippets. In: Proceedings of the 23rd International Conference on World Wide Web. WWW '14 Companion, New York, NY, USA, ACM (2014) 33–38

[7] De Vries, A.P., Vercoustre, A.M., Thom, J.A., Craswell, N., Lalmas, M.: Overview of the inex 2007 entity ranking track. In: International Workshop of the Initiative for the Evaluation of XML Retrieval, Springer (2007) 245–251

[8] Gabrilovich, E., Ringgaard, M., Subramanya, A.: Facc1: Freebase annotation of clueweb corpora, version 1 (release date 2013-06-26, format version 1, correction level 0). Note: http://lemurproject.org/clueweb09/FACC1/Cited by **5** (2013)

[9] Gillick, D., Favre, B.: A scalable global model for summarization. In: Proceedings of the Workshop on Integer Linear Programming for Natural Langauge Processing. ILP '09, Stroudsburg, PA, USA, Association for Computational Linguistics (2009) 10–18

[10] Goldstein, J., Mittal, V., Carbonell, J., Kantrowitz, M.: Multi-document summarization by sentence extraction. In: Proceedings of the 2000 NAACL-ANLPWorkshop on Automatic Summarization - Volume 4. NAACL-ANLP-AutoSum '00, Stroudsburg, PA, USA, Association for Computational Linguistics (2000) 40–48

[11] Gong, S., Qu, Y., Tian, S.: Summarization using wikipedia. In: TAC. (2010)

[12] Gunaratna, K., Thirunarayan, K., Sheth, A.P.: FACES: diversity-aware entity summarization using incremental hierarchical conceptual clustering. In Bonet, B., Koenig, S., eds.: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA., AAAI Press (2015) 116–122

[13] Herbrich, R., Minka, T., Graepel, T.: Trueskill$^{\text{TM}}$: A bayesian skill rating system. In: Proceedings of the 19th International Conference on Neural Information Processing Systems. NIPS'06, Cambridge, MA, USA, MIT Press (2006) 569–576

[14] Hoffmann, R., Zhang, C., Ling, X., Zettlemoyer, L., Weld, D.S.: Knowledge-based weak supervision for information extraction of overlapping relations. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1. HLT '11, Stroudsburg, PA, USA, Association for Computational Linguistics (2011) 541–550

[15] Ji, H., Grishman, R., Dang, H.T., Griffitt, K., Ellis, J.: Overview of the tac 2010 knowledge base population track. In: Third Text Analysis Conference (TAC 2010). Volume 3. (2010) 3–13

[16] Larocca Neto, J., Santos, A.D., Kaestner, C.A.A., Freitas, A.A.: Document clustering and text summarization. In Mackin, N., ed.: Proceedings of the 4th International Conference Practical Applications of Knowledge Discovery and Data Mining (PADD-2000), London, The Practical Application Company (January 2000) 41–55

[17] Mason, R., Gaska, B., Durme, B.V., Choudhury, P., Hart, T., Dolan, B., Toutanova, K., Mitchell, M.: Microsummarization of online reviews: An experimental study. In: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA. (2016) 3015–3021

[18] McDonald, R.: A study of global inference algorithms in multi-document summarization. In: European Conference on Information Retrieval, Springer (2007) 557–564

[19] McNamee, P., Dang, H.T.: Overview of the tac 2009 knowledge base population track. In: Text Analysis Conference (TAC). Volume 17. (2009) 111–113

[20] Meng, X., Wei, F., Liu, X., Zhou, M., Li, S., Wang, H.: Entity-centric topic-oriented opinion summarization in twitter. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '12, New York, NY, USA, ACM (2012) 379–387

[21] Mintz, M., Bills, S., Snow, R., Jurafsky, D.: Distant supervision for relation extraction without labeled data. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2. ACL '09, Stroudsburg, PA, USA, Association for Computational Linguistics (2009) 1003–1011

[22] Nenkova, A., Passonneau, R., McKeown, K.: The pyramid method: Incorporating human content selection variation in summarization evaluation. ACM Trans. Speech Lang. Process. **4**(2) (May 2007)

[23] Pradhan, S., Ramshaw, L., Marcus, M., Palmer, M., Weischedel, R., Xue, N.: Conll-2011 shared task: Modeling unrestricted coreference in ontonotes. In: Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task, Association for Computational Linguistics (2011) 1–27

[24] Riedel, S., Yao, L., McCallum, A., Marlin, B.M.: Relation extraction with matrix factorization and universal schemas. In Vanderwende, L., III, H.D., Kirchhoff, K., eds.: HLT-NAACL, The Association for Computational Linguistics (2013) 74–84

[25] Sakaguchi, K., Post, M., Durme, B.V.: Efficient elicitation of annotations for human evaluation of machine translation. In: Proceedings of the Ninth Workshop on Statistical Machine Translation, WMT@ACL 2014, June 26-27, 2014, Baltimore, Maryland, USA, The Association for Computer Linguistics (2014) 1–11

[26] Salton, G., Allan, J., Buckley, C.: Approaches to passage retrieval in full text information systems. In: Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '93, New York, NY, USA, ACM (1993) 49–58

[27] Sundheim, B.M.: Overview of results of the muc-6 evaluation. In: Proceedings of a workshop on held at Vienna, Virginia: May 6-8, 1996, Association for Computational Linguistics (1996) 423–442

[28] Surdeanu, M., Tibshirani, J., Nallapati, R., Manning, C.D.: Multi-instance multi-label learning for relation extraction. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. EMNLP-CoNLL '12, Stroudsburg, PA, USA, Association for Computational Linguistics (2012) 455–465

[29] Thalhammer, A., Knuth, M., Sack, H. In: Evaluating Entity Summarization Using a Game-Based Ground Truth. Springer Berlin Heidelberg, Berlin, Heidelberg (2012) 350–361

[30] Verga, P., Neelakantan, A., McCallum, A.: Generalizing to unseen entities and entity pairs with row-less universal schema. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, Valencia, Spain, Association for Computational Linguistics (April 2017) 613–622