Moving away from semantic overfitting in disambiguation datasets

Marten Postma and **Filip Ilievski** and **Piek Vossen** and **Marieke van Erp**
Vrije Universiteit Amsterdam
m.c.postma,f.ilievski,piek.vossen,marieke.van.erp@vu.nl

Abstract

Entities and events in the world have no frequency, but our communication about them and the expressions we use to refer to them do have a strong frequency profile. Language expressions and their meanings follow a Zipfian distribution, featuring a small amount of very frequent observations and a very long tail of low frequent observations. Since our NLP datasets sample texts but do not sample the world, they are no exception to Zipf's law. This causes a lack of representativeness in our NLP tasks, leading to models that can capture the head phenomena in language, but fail when dealing with the long tail. We therefore propose a referential challenge for semantic NLP that reflects a higher degree of ambiguity and variance and captures a large range of small real-world phenomena. To perform well, systems would have to show deep understanding on the linguistic tail.

1 Introduction

Semantic processing addresses the relation between natural language and a representation of a world, to which language makes reference. A challenging property of this relation is the context-bound complex interaction between lexical expressions and world meanings.[1] Like many natural phenomena, the distribution of expressions and their meanings follows a power law such as Zipf's law (Newman, 2005) , with a few very frequent observations and a very long tail of low frequent observations.[2] Still, the world itself has no frequency. All entities and events in the world appear to us with a frequency of 1. Nevertheless, we dominantly talk about only a few instances in the world and refer to them with a small set of expressions, which can only be explained by the contextual constraints within a language community, a topic, a location, and a period of time. Without taking these into account, it is impossible to fully determine meaning.

Given that instances in the world do not have frequencies, language and our writing about the world is heavily skewed, selective, and biased with respect to that world. A name such as *Ronaldo* can have an infinite amount of references and in any world (real or imaginary) each *Ronaldo* is equally present. Our datasets, however, usually make reference to only one *Ronaldo*. The problem, as we see it, is that our NLP datasets sample texts but do not sample the world. This causes lack of representativeness in our NLP tasks, that has big consequences for language models: they tend to capture the head phenomena in text without considering the context constraints and thus fail when dealing with less dominant world phenomena. As a result, there is little awareness of the full complexity of the task in relation to the contextual realities, given language as a system of expressions and the possible interpretations within contexts of time, location, community, and topic. People, however, have no problem to handle local real-world situations that are referenced to in text.

We believe it is time to create a task that encour-

[1] We use *meaning* as an umbrella term for both concepts or lexical meanings and instances or entities, and *lexical expression* as a common term for both lemmas and surface forms.

[2] We acknowledge that there also exist many long tail phenomena in syntactic processing, e.g. syntactic parsing.

Proceedings of EMNLP 2016 Workshop on Uphill Battles in Language Processing: Scaling Early Achievements to Robust Methods, pages 17–21,
Austin, TX, November 5, 2016.

ages systems to model the full complexity of disambiguation by enriched context awareness. We hence propose a semantic referential challenge, event-based Question Answering (QA), that reflects a high degree of ambiguity and variance and captures a wide range of small real-world phenomena. This task requires a deeper semantic understanding of the linguistic tail of several disambiguation challenges.

2 Related work

We present related work on the representativeness of the disambiguation datasets (Section 2.1), as well as the representativeness of the QA task (Section 2.2).

2.1 The Long Tail in disambiguation tasks

Past work tried to improve the disambiguation complexity. Vossen et al. (2013) created a balanced corpus, DutchSemCor, for the Word Sense Disambiguation (WSD) task in which each sense gets an equal number of examples. Guha et al. (2015) created the QuizBowl dataset for Entity Coreference (EnC), while Cybulska and Vossen (2014) extended the existing Event Coreference (EvC) dataset ECB to ECB+, both efforts resulting in notably greater ambiguity and temporal diversity. Although all these datasets increase the complexity for disambiguation, they still contain a limited amount of data which is far from approximating realistic tasks.

Properties of existing disambiguation datasets have been examined for individual tasks. For WSD, the correct sense of a lemma is shown to often coincide with the most frequent sense (Preiss, 2006). Van Erp et al. (2016) conclude that Entity Linking (EL) datasets contain very little referential ambiguity and focus on well-known entities, i.e. entities with high PageRank (Page et al., 1999) values. Moreover, the authors note a considerable overlap of entities across datasets. Cybulska and Vossen (2014) and Guha et al. (2015) both stress the low ambiguity in the current datasets for the tasks of EvC and EnC.

In Ilievski et al. (2016), we measure the properties of existing disambiguation datasets for the tasks of EL, WSD, EvC, EnC, and Semantic Role Labeling (SRL), through a set of generic representation metrics applicable over tasks. The analyzed datasets show a notable bias with respect to aspects of ambiguity, variance, dominance, and time, thus exposing a strong semantic overfitting to a very limited, and within that, popular part of the world.

The problem of overfitting to a limited set of test data has been addressed by the field of domain adaptation (Daume III, 2007; Carpuat et al., 2013; Jiang and Zhai, 2007). In addition, unsupervised domain-adversarial approaches attempt to build systems that generalize beyond the specifics of a given dataset, e.g. by favoring features that apply to both the source and target domains (Ganin et al., 2016). By evaluating on another domain than the training one, these efforts have provided valuable insights into system performance. Nevertheless, this research has not addressed the aspects of time and location. Moreover, to our knowledge, no approach has been proposed to generalize the problem of reference to unseen domains, which may be due to the enormous amount of references that exist in the world leading to an almost infinite amount of possible classes.

2.2 The Long Tail in QA tasks

The sentence selection datasets WikiQA (Yang et al., 2015) and QASent (Wang et al., 2007) consist of questions that are collected from validated user query logs, while the answers are annotated manually from automatically selected Wikipedia pages. WIKIREADING (Hewlett et al., 2016) is a recent large-scale dataset that is based on the structured information from Wikidata (Vrandečić and Krötzsch, 2014) and the unstructured information from Wikipedia. Following a smart fully-automated data acquisition strategy, this dataset contains questions about 884 properties of 4.7 million instances. While these datasets require semantic text processing of the questions and the candidate answers, there is a finite set of answers, many of which represent popular interpretations from the world, as a direct consequence of using Wikipedia. To our knowledge, no QA task has been created to deliberately address the problem of (co)reference to long tail instances, where the list of potential interpretations is enormous, largely ambiguous, and only relevant within a specific context. The long tail aspect could be emphasized by an event-driven QA task, since the referential ambiguity of events in the world is much higher than the ambiguity of entities. No event-driven QA task has been proposed in past work. As Wikipedia only represents a tiny and popular subset

of all world events, the Wikipedia-based approaches could not be applied to create such a task, thus signaling the need for a novel data acquisition approach to create an event-driven QA task for the long tail.

Weston et al. (2015) propose 20 skill sets for a comprehensive QA system. This work presents reasoning categories (e.g. spatial and temporal reasoning) and requires within-document coreference, which are very relevant skills for understanding linguistic phenomena of the long tail. However, the answer in these tasks is usually mentioned in the text, thus not addressing the referential complexity of the long tail phenomena in the world.

3 Moving away from semantic overfitting

Current datasets only cover a small portion of the full complexity of the disambiguation task, focusing mostly on the head. This has encouraged systems to overfit on the head and largely ignore the linguistic tail. Due to this lack of representativeness, we are not able to determine to which degree systems achieve language understanding of the long tail.

As described in the previous Section, the challenge of semantic overfitting has been recognized by past work. QA datasets, such as WIKIREADING, have increased the complexity of interpretation by using a large number of entities and questions, also allowing for subsets to be sampled to tackle specific tasks. The skill sets presented by Weston et al. (2015) include long tail skills that are crucial in order to interpret language in various micro-contexts. None of these approaches has yet created a task that addresses the long tail explicitly and recognizes the full referential complexity of disambiguation. Considering the competitiveness of the field, such task is necessary to motivate systems that can deal with the long tail and adapt to new contexts.

We therefore advocate a task that requires a deep semantic processing linked to both the head and the long tail. It is time to create a high-level referential challenge for semantic NLP that reflects a higher degree of ambiguity and variation and captures a wide range of small real-world phenomena. This task can not be solved by only capturing the head phenomena of the disambiguation tasks in any sample text collection. For maximum complexity, we propose an event-driven QA task that also represents local events, thus capturing phenomena from both the head and the long tail. Also, these events should be described across multiple documents that exhibit a natural topical spread over time, providing information bit-by-bit as it becomes available.

4 Task requirements

We define five requirements that should be satisfied by an event-driven QA task in order to maximize confusability, to challenge systems to deal with the tail of the Zipfian distribution, and to adapt to new contexts. These requirements apply to a single event topic, e.g. *murder*. Each event topic should contain:

R1 Multiple event instances per event topic, e.g. *the murder of John Doe* and *the murder of Jane Roe*.

R2 Multiple event mentions per event instance within the same document.

R3 Multiple documents with varying document creation times in which the same event instances are described to capture topical information over time.

R4 Event confusability by combining one or multiple confusion factors:

a) ambiguity of event surface forms, e.g. *John Smith fires a gun*, and *John Smith fires an employee*.

b) variance of event surface forms, e.g. *John Smith kills John Doe*, and *John Smith murders John Doe*.

c) time, e.g. *murder A that happened in January 1993*, and *murder B in October 2014*.

d) participants, e.g. *murder A committed by John Doe*, and *murder B committed by the Roe couple*.

e) location, e.g. *murder A that happened in Oklahoma*, and *murder B in Zaire*.

R5 Representation of non-dominant events and entities, i.e. instances that receive little media coverage. Hence, the entities would not be restricted to celebrities and the events not to general elections.

5 Proposal

We propose a semantic task that represents the linguistic long tail. The task will consist of one high-level challenge (QA), for which an understanding of the long tail of several disambiguation tasks (EL, WSD, EvC, EnC) is needed in order to perform well on the high-level challenge. The QA task would feature two levels of event-oriented questions: instance-level questions (e.g. *Who was killed*

last summer in Vienna?) and aggregation-level questions (e.g. *How many white people have been poisoned in the last 2 years?*). The setup would be such that the QA challenge could in theory be addressed without performing any disambiguation (e.g. using enhanced Information Retrieval), but deeper processing, especially on the disambiguation tasks, would be almost necessary in practice to be able to come up with the correct answers.

To some extent, the requirements in Section 4 are satisfied by an existing corpus, ECB+ (Cybulska and Vossen, 2014), which contains 43 event topics. For each event topic in the corpus, there are at least 2 different seminal events (R1). Since the corpus contains 7,671 intra-document coreference links, on average 7.8 per document, we can assume that requirement R2 is satisfied to a large extent. Although there are multiple news articles per event instance, they are not spread over time, which means that R3 is not satisfied. Furthermore, the event confusability factors (R4) are not fully represented, since the ambiguity and variance of the event surface forms and the participants are still very low, whereas the dominance is quite standard (R5), which is not surprising given that these aspects were not considered during the corpus assembly period. Additionally, only 1.8 sentences per document were annotated on average. Potential references in the remaining sentences need to be validated as well.

We will start with the ECB+ corpus and expand it by following an event topic-based strategy:[3]

1) Pick a subset of ECB+ topics, by favoring: a) seminal events (e.g. *murder*) whose surface forms have a low lexical ambiguity, but can be referred to by many different surface forms (*execute, slay, kill*)
b) combinations of two or more seminal events that can be referred to by the same polysemous form (e.g. *firing*).

2) Select one or more confusability factors from R4, e.g. by choosing *participants* and *variance*. This step can be repeated for different combinations of the confusability factors.

3) Increase the amount of events for an event topic (to satisfy R1). We add new events based on the confusability factors chosen in step 2 and from local news sources to ensure low dominance (R5). These events can come from different documents or the same document.

4) Retrieve multiple event mentions for each event based on the decision from the confusability factors (R4). We use local news sources to ensure low dominance (R5). They originate from the same document (R2) and from different documents with (slightly) different creation times (R3).

In order to facilitate the expansion described in steps 3 and 4, we will add documents to ECB+ from The Signal Media One-Million News Articles Dataset (Signal1M) (Corney et al., 2016). This will assist in satisfying requirement R5, since the Signal1M Dataset is a collection of mostly local news. For the expansion, active learning will be applied on the Signal1M Dataset, guided by the decisions in step 2, to decide which event mentions are coreferential and which are not. By following our four-step acquisition strategy and by using the active learning method, we expect to obtain a high accuracy on EvC. As we do not expect perfect accuracy of EvC even within this smart acquisition, we will validate the active learning output. The validation will lead to a reliable set of events on a semantic level for which we would be able to pose both instance-level and aggregation-level questions, as anticipated earlier in this Section. As the task of QA does not require full annotation of all disambiguation tasks, we would be able to avoid excessive annotation work.

6 Conclusions

This paper addressed the issue of semantic overfitting to disambiguation datasets. Existing disambiguation datasets expose lack of representativeness and bias towards the head interpretation, while largely ignoring the rich set of long tail phenomena. Systems are discouraged to consider the full complexity of the disambiguation task, since the main incentive lies in modelling the head phenomena. To address this issue, we defined a set of requirements that should be satisfied by a semantic task in order to inspire systems that can deal with the linguistic tail and adapt to new contexts. Based on these requirements, we proposed a high-level task, QA, that requires a deep understanding of each disambiguation task in order to perform well.

[3]The same procedure can be followed for an entity-centric expansion approach.

References

[Carpuat et al.2013] Marine Carpuat, Hal Daumé III, Katharine Henry, Ann Irvine, Jagadeesh Jagarlamudi, and Rachel Rudinger. 2013. Sensespotting: Never let your parallel data tie you to an old domain. In *Proceedings of the Association for Computational Linguistics (ACL)*. Citeseer.

[Corney et al.2016] David Corney, Dyaa Albakour, Miguel Martinez, and Samir Moussa. 2016. What do a million news articles look like? In *Proceedings of the First International Workshop on Recent Trends in News Information Retrieval co-located with 38th European Conference on Information Retrieval (ECIR 2016), Padua, Italy, March 20, 2016.*, pages 42–47.

[Cybulska and Vossen2014] Agata Cybulska and Piek Vossen. 2014. Using a sledgehammer to crack a nut? lexical diversity and event coreference resolution. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pages 4545–4552.

[Daume III2007] Hal Daume III. 2007. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL 2007)*, pages 256–263, Prague, Czech Republic. Association for Computational Linguistics.

[Ganin et al.2016] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35.

[Guha et al.2015] Anupam Guha, Mohit Iyyer, Danny Bouman, Jordan Boyd-Graber, and Jordan Boyd. 2015. Removing the training wheels: A coreference dataset that entertains humans and challenges computers. In *North American Association for Computational Linguistics (NAACL)*.

[Hewlett et al.2016] Daniel Hewlett, Alexandre Lacoste, Llion Jones, Illia Polosukhin, Andrew Fandrianto, Jay Han, Matthew Kelcey, and David Berthelot. 2016. Wikireading: A novel large-scale language understanding task over wikipedia. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1535–1545. Association for Computational Linguistics.

[Ilievski et al.2016] Filip Ilievski, Marten Postma, and Piek Vossen. 2016. Semantic overfitting: what 'world' do we consider when evaluating disambiguation of text? *In proceedings of COLING*.

[Jiang and Zhai2007] Jing Jiang and ChengXiang Zhai. 2007. Instance weighting for domain adaptation in NLP. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL 2007)*, volume 7, pages 264–271.

[Newman2005] Mark EJ Newman. 2005. Power laws, pareto distributions and zipf's law. *Contemporary physics*, 46(5):323–351.

[Page et al.1999] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The PageRank citation ranking: bringing order to the web.

[Preiss2006] Judita Preiss. 2006. A detailed comparison of WSD systems: an analysis of the system answers for the Senseval-2 English all words task. *Natural Language Engineering*, 12(03):209–228.

[van Erp et al.2016] Marieke van Erp, Pablo Mendes, Heiko Paulheim, Filip Ilievski, Julien Plu, Giuseppe Rizzo, and Jorg Waiterlonis. 2016. Evaluating entity linking: An analysis of current benchmark datasets and a roadmap for doing a better job. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*.

[Vossen et al.2013] Piek Vossen, Ruben Izquierdo, and Atilla Görög. 2013. DutchSemCor: in quest of the ideal sense-tagged corpus. In *Proceedings of Recent Advances in Natural Language Processing (RANLP)*, pages 710–718. INCOMA Ltd. Shoumen, Bulgaria.

[Vrandečić and Krötzsch2014] Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

[Wang et al.2007] Mengqiu Wang, Noah A Smith, and Teruko Mitamura. 2007. What is the jeopardy model? a quasi-synchronous grammar for qa. In *EMNLP-CoNLL*, volume 7, pages 22–32.

[Weston et al.2015] Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart van Merriënboer, Armand Joulin, and Tomas Mikolov. 2015. Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*.

[Yang et al.2015] Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. Wikiqa: A challenge dataset for open-domain question answering. In *Proceedings of EMNLP*, pages 2013–2018. Citeseer.