Semantic overfitting: what 'world' do we consider when evaluating disambiguation of text?

Filip Ilievski, Marten Postma, Piek Vossen
Vrije Universiteit Amsterdam
De Boelelaan 1105, 1081 HV Amsterdam
The Netherlands
{f.ilievski, m.c.postma, piek.vossen}@vu.nl

Abstract

Semantic text processing faces the challenge of defining the relation between lexical expressions and the world to which they make reference within a period of time. It is unclear whether the current test sets used to evaluate disambiguation tasks are representative for the full complexity considering this time-anchored relation, resulting in semantic overfitting to a specific period and the frequent phenomena within. We conceptualize and formalize a set of metrics which evaluate this complexity of datasets. We provide evidence for their applicability on five different disambiguation tasks. To challenge semantic overfitting of disambiguation systems, we propose a time-based, metric-aware method for developing datasets in a systematic and semi-automated manner, as well as an event-based QA task.

1 Introduction

Semantic processing defines a relation between natural language and a representation of a world it refers to. A challenging property of natural language is the time-bound complex interaction between lexical expressions and world meanings. We use *meaning* in this paper as an umbrella term for both concepts and (event and entity) instances, and *lexical expression* as a common term for both lemmas and surface forms. We can define this interaction as a set of relations, both sense relations and referential relations, that exists within a language community in a certain period of time, e.g. one or a few generations. The people belonging to these generations share one language system that changes relatively slowly but during their lives there are many rapidly changing situations in the world that make certain meanings and expressions dominant and others not. Likewise, we expect that a generation uses a certain set of lexical expressions out of the available set in relation to a set of meanings that balances the trade-off between learning many expressions and resolving extreme ambiguity of a small set of expressions.

The task of interpreting lexical expressions as meanings, known as disambiguation, has been addressed by the NLP community following a "divide & conquer" strategy that mostly ignores this complex time-bound relation. Over the years, this resulted in numerous separate disambiguation tasks each with a specific set of datasets restricted to a small bandwidth with respect to the dynamics of the world and the large scope of the possible meanings that lexical expressions can have. By dividing the problem into different tasks on relatively small datasets, researchers can focus on specific subproblems and have their efforts evaluated in a straightforward manner. Datasets have been developed independently for each task, intended as a test bench to evaluate the accuracy and applicability of the proposed systems. Official evaluation scripts have been created for most datasets to enable a fair comparison across systems.

The downside of this practice is that task integration is discouraged, systems tend to be optimized on the few datasets available for each task, and the dependencies of ambiguities across tasks in relation to the time-bound contextual realities are not considered. As a result, there is little awareness of the overall complexity of the task, given language as a system of expressions and the possible interpretations given the changing world over longer periods of time. Systems are thus encouraged to strongly overfit on a single task, a single dataset, and a specific 'piece' of the world at a specific moment in time.

Each text forms a unique semantic puzzle of expressions and meanings in which ambiguity is limited within the specific time-bound context, but is extreme without considering this context. The main question we thus put forward and address in this paper is how to enhance disambiguation tasks to cover the full complexity of the time-bound interaction between lexical expressions and meanings (in the broad sense of the word as defined here). We therefore first propose a number of metrics that formally quantity the complexity of this relation and apply this to a wide range of available datasets for a broad range of semantic tasks. Secondly, we provide evidence for the limitations of the current tasks and, thirdly, we present a proposal to improve these tasks in the hope that we challenge future research to address these limitations.

The paper is structured as follows. We motivate the importance and relevance of this temporal interaction for both concept- and instance-based disambiguation tasks in Section 2. Following up on previous research (Section 3), we define a model of the complex interaction (Section 4), and we conceptualize and formalize a collection of metrics in a generic manner (Section 5). Moreover, we apply these metrics to quantify aspects of existing evaluation sets (Section 6). In Section 7, we propose two approaches for creating metric-aware test sets that include a temporal dimension. The paper is concluded in Section 8.

2 Temporal Aspect of the Disambiguation Task

We live in a dynamic and rapidly changing world: some companies expand their offices all around the globe, while others collapse; people become celebrities overnight and are forgotten only several years afterwards. Similarly, a whole range of mainstream technological concepts of today's world have only been known since the last few decades. These observations have a big impact on the dynamics of a language system, since the relation between language expressions and meanings follows the changes in the world. To some extent this is reflected in new expressions and new meanings but most strongly this is reflected in the distributional usage of expressions and their dominant meaning.

For instance, the dominant meaning of the terms *mobile*, *cell*, and *phone* is the same for the contemporary, especially young, generations: mobile phone. On the other hand, older generations also remember different dominant concepts from the 80s and 90s: *mobile* being typically a decoration hanging from the ceiling, *cell* usually being a unit in a prison or body tissue, while *phone* referring to the static devices found at home or on the streets. The dominant meanings of the 80s and 90s have been replaced by new dominant meanings, whereas the younger generation may have lost certain meanings such as the decoration. Similarly, football fans remember two different superstar *Ronaldo* players which have been dominant one after the other: the Brazillian striker and the Portuguese Ballon d'Or award winner.

What is shown by these examples is that not only new meanings appear and old meanings become obsolete but that, more strongly, the usage distribution of competing meanings changes over time. As the mobile phone gains popularity and the mobile decoration gets replaced by others, people refer to the mobile phone more often than the traditional mobile decoration. Hence, in a later point of time, the most commonly used meaning for *mobile* changes, even though both meanings are still possible. Similarly for the *Ronaldo* case: in 2016 one can still refer to both players, but the dominant meaning is now the Portuguese player.

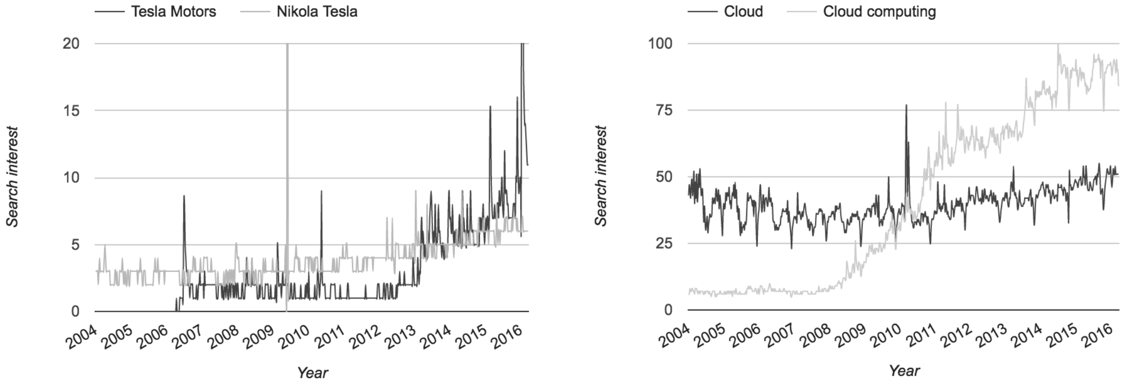
We also observe a relation between the variety of lexical expressions used to refer to a meaning, and its dominance of usage. As the mobile phone gained popularity, its set of associated expressions expanded from only *mobile phone* to also: *mobile*, *phone*, *cell phone*, and *cell*. On the other hand, when referring to a prison cell without a specific context, one should nowadays explicitly use the full expression *prison cell* instead of just *cell*.

To measure the usage distribution of competing meanings, we could use online resources that track these distributions over time, such as Google Trends[1] and Wikipedia Views.[2] We present the usage distribution for instances denoted by *Tesla* in Figure 1a, and for concepts expressed with the expression *cloud* in Figure 1b. These plots demonstrate the ways in which the distribution of usage changes both for instances and concepts as a function of the temporal dimension. As we discussed in Section 1, the notion

[1] https://www.google.com/trends/

[2] http://stats.grok.se/

Figure 1: Usage distribution for ambiguous concepts and instances based on Google Trends data.



(a) Usage of two dominant meanings for *Tesla*. The black line depicts the usage of *Tesla Motors Company*, while the gray line represents *Nikola Tesla*.

(b) Usage of two dominant meanings for *cloud*. The black line depicts the usage of the clouds in the sky, natural objects placed in the atmosphere. The gray line stands for the modern meaning of cloud as an Internet-based type of computing.

of time and its role in this mapping between expressions and meanings has not been taken into account in the creation of existing disambiguation datasets. This observation points to a serious weakness in the representativeness of existing datasets for the full complexity of the disambiguation task. Consequently, systems are not encouraged to focus on the temporal aspect of the task but in reality the same language system is still used for many different situations within a changing world. While this works for humans, this is not yet solved for machines.

3 Related Work

The three problems enumerated in Section 1 have been addressed to some extent in past work.

Several approaches have attempted to resolve pairs of disambiguation tasks jointly. Examples include: combined Entity Linking (EL) and Word Sense Disambiguation (WSD) (Hulpuş et al., 2015; Moro et al., 2014), combined event and entity coreference (EvC and EnC) (Lee et al., 2012) and resolving WSD and Semantic Role Labeling (SRL) together (Che and Liu, 2010). Although some task combinations are well-supported by multi-task datasets, such as CoNLL 2011 and 2012 for joint coreference (Pradhan et al., 2011; Pradhan et al., 2012), and Moro and Navigli (2015) for WSD and EL, still many multi-task systems have to be evaluated on separate datasets. Notable efforts to create multi-task annotated corpora are the AMR Bank (Banarescu et al., 2013) and the MEANTIME corpus (Minard et al., 2016a).

Properties of existing 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) or the predominant sense (McCarthy et al., 2004). In the case of McCarthy et al. (2004), the predominant sense is deliberately adapted with respect to the topic of the text. Our work differs from McCarthy et al. (2004) because they do not consider the temporal dimension. As a response to sense-skewed datasets, Vossen et al. (2013) created a balanced sense corpus in the DutchSemCor project in which each sense gets an equal number of examples. Similarly, Van Erp et al. (2016) conclude that EL datasets contain very little referential ambiguity. Evaluation is focused on well-known entities, i.e. entities with high PageRank (Page et al., 1999) values. Additionally, the authors observe a considerable overlap of entities across datasets, even for pairs of datasets that represent entirely different topics. 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, respectively. Motivated by these findings, Guha et al. (2015) created a new dataset (QuizBowl), while Cybulska and Vossen (2014) extended the existing dataset ECB to ECB+, both efforts resulting in notably greater ambiguity and temporal diversity. As far as we are aware, no existing disambiguation dataset has included the temporal dependency of ambiguity, variance, or dominance.

The problem of overfitting to a limited set of test data has been of central interest to the body of work focusing on domain adaptation (Daume III, 2007; Carpuat et al., 2013; Jiang and Zhai, 2007). By evaluating on a different domain than the training one, these efforts have provided valuable insights into system performance. However, to our knowledge, this research has also not addressed the temporal aspect of the task.

We therefore propose to take this a step further and examine system performance with respect to a set of metrics, applicable over disambiguation tasks, thus setting the stage for creation of metric-aware datasets. We expect that these metrics show reduced complexity within well-defined temporal and topical boundaries and increased complexity across these boundaries. More extensive datasets than existing single- and multi-task datasets, driven by metrics on ambiguity, variance, dominance and time, would challenge semantic overfitting.

4 Semiotic Generation and Context Model

We want to model the relation between expressions and meanings in the world within a generation that shares the same language system, as well as the fluctuation in usage of expressions and meanings over time within this generation. We therefore assume that for each language community at a specific time, there exist a set of meanings M in the world and a set of lexical expressions L in a language. The relation between these sets is many-to-many: each lexical expression L_i can refer to multiple meanings $M_1, M_2, ...$ (ambiguity) and each meaning M_j can be verbalized through multiple lexical expressions $L_1, L_2, ...$ (variance). As we discuss in Section 2, the sets of M, L, their relations, and especially the distributions of these relations, are dynamic, i.e. they can change over time. We denominate this model "Semiotic Generation and Context Model", because it captures the distribution changes in the semiotic relation between meanings and lexical expressions, given the context of the changes in the world and within the language system of a generation.

In practice, we study available proxies of the world at a moment in time and of the language of a generation which capture this relation at a given time snapshot: lexical resources are considered as a proxy of the language system of a generation and the dataset is considered as a proxy for the world at a particular moment in time creating a specific context. We analyze the time-anchored interaction between M and L in the datasets proxy and measure this against their interaction in the resources proxy to provide insight on how representative the datasets are for the task. Note that the proxies of datasets and resources cover only a subset of the language used within a generation, and (consequently) only a subset of all possible meanings. While not ideal, this is the best we have because there is no way to capture all language used within a generation nor possibly list every possible meaning, especially considering that we can always create new meanings, e.g. by inventing some non-real world ones.

5 Methodology

Based on the Semiotic Generation and Context Model, we now define and formalize a number of metrics that qualify datasets for disambiguation tasks. In this Section, we describe these metrics and explain the tasks we focus on. Furthermore, we enumerate the design choices that guide our pick of datasets and we elaborate on the datasets we analyze.

5.1 Metrics

Mean Observed Ambiguity (MOA)

We define observed ambiguity of an expression as the cardinality of the set of meanings it refers to within a dataset (O_{L_i}). For example, the expression *horse* has 4 meanings in WordNet but only the chess meaning occurs in the dataset, resulting in an observed ambiguity of 1. The Mean Observed Ambiguity (MOA) of a dataset is then the average of the individual observed ambiguity values.

Mean Observed Variance (MOV)

We define observed variance of a meaning as the cardinality of the set of lexical expressions that express it within a dataset (O_{M_j}). The chess meaning of *horse* also has *knight* as a synonym but only *horse* occurs in the dataset, hence an observed variation of 1. The Mean Observed Variance (MOV) of a dataset is then the average of the individual observed variance values.

Mean Observed Dominance of Ambiguity (MODA)
We define dominance of ambiguity as a frequency distribution of the dominant meaning of a lexical expression. For example, *horse* occurs 100 times in the data and in 80 cases it has the chess meaning: the dominance score is 0.8. The Mean Observed Dominance of Ambiguity (MODA) of a dataset is the average dominance of all observed expressions.

Mean Observed Dominance of Variance (MODV)
We define the notion of dominance of variance, as a frequency distribution of the dominant lexical expression referring to a meaning. If *horse* is used 60 times and *knight* 40 times for the same meaning then the observed dominance of variance is 0.6. The Mean Observed Dominance of Variance (MODV) of a dataset is then the average dominance computed over all observed meanings.

Entropy of the Meanings (Normalized) of a Lexical Expression (EMNLE)
We define an alternative notion of dominance, based on entropy, in order to consider the distribution of the less dominant classes in a dataset. We introduce $p(M_j|L_i)$: a conditional probability of a meaning M_j based on the occurrence of a lexical expression L_i. We compute this probability using the formula $p(M_j|L_i) = \frac{p(M_j,L_i)}{p(L_i)}$, a ratio between the number of common occurrences of M_j and L_i, and on the other hand, occurrences of L_i alone. We combine the individual conditional probabilities for L_i in a single information theory metric of entropy, $H(O_{L_i})$:

$$H(O_{L_i}) = \frac{-\sum\limits_{j=1}^{n} p(M_j|L_i) log_2 p(M_j|L_i)}{log_2(n)} \tag{1}$$

For example, given 100 occurrences of the lexical expression *horse*, where 80 occurrences refer to the the chess meaning and 20 to the animal meaning, the entropy of the expression *horse* would be 0.72. To compute a single entropy (EMNLE) value over all lexical expressions in a dataset, we average over the individual entropy values:

$$EMNLE(O_L, R_L) = \frac{1}{n}\sum_{i=1}^{n} H(O_{L_i}, R_{L_i}) \tag{2}$$

Entropy of the Lexical Expressions (Normalized) of a Meaning (ELENM)
We introduce $p(L_i|M_j)$: a conditional probability of a lexical expression L_i based on the occurrence of a meaning M_j. We compute this probability using the formula $p(L_i|M_j) = \frac{p(L_i,M_j)}{p(M_j)}$, a ratio between the number of common occurrences of M_j and L_i, and on the other hand, occurrences of M_j alone. We combine the individual conditional probabilities for M_j in a single information theory metric of entropy, $H(O_{M_j})$:

$$H(O_{M_j}) = \frac{-\sum\limits_{i=1}^{n} p(L_i|M_j) log_2 p(L_i|M_j)}{log_2(n)} \tag{3}$$

Suppose the meaning of horse as a chess piece is expressed 60 times by the lexical expression *horse* and 40 times by *knight*, then the entropy of the chess piece meaning of *horse* is 0.97. To compute a single entropy (ELENM) value over all meanings in a dataset, we average over the individual entropy values:

$$ELENM(O_M, R_M) = \frac{1}{n}\sum_{j=1}^{n} H(O_{M_j}, R_{M_j}) \tag{4}$$

Relation between Observed and Resource Ambiguity (RORA)
We define resource ambiguity of a lexical expression as the cardinality of the set of meanings that it can refer to according to a lexical resource (R_{L_i}). Then we define the ratio between observed and resource ambiguity for a lexical expression as:

$$ratio_{amb}(O_{L_i}, R_{L_i}) = \frac{|\{M_j : M_j \in O_{L_i}\}|}{|\{M_j : M_j \in R_{L_i}\}|} \tag{5}$$

In the case that only 1 out of 4 resource meanings is observed in the dataset, for example only the chess meaning of *horse*, this would lead to a $ratio_{amb}$ value of 0.25. To compute the RORA value of a dataset, we average over the individual ratios:

$$RORA(O_L, R_L) = \frac{1}{n}\sum_{i=1}^{n} ratio_{amb}(O_{L_i}, R_{L_i}) \tag{6}$$

Relation between Observed and Resource Variance (RORV)

We define resource variance of a meaning as the cardinality of the set of lexical expressions which can verbalize it (R_{M_j}). Then we define the ratio between observed and resource variance for a given meaning:

$$ratio_{var}(O_{M_j}, R_{M_j}) = \frac{|\{L_i : L_i \in O_{M_j}\}|}{|\{L_i : L_i \in R_{M_j}\}|} \tag{7}$$

Suppose that the expressions *horse* and *knight* can refer to the meaning of chess piece according to a resource, but only the expression *horse* refers to it in a particular dataset, this would lead to a $ratio_{var}$ value of 0.5. To compute the RORV value of a dataset, we average over the individual ratios:

$$RORV(O_M, R_M) = \frac{1}{n}\sum_{i=1}^{n} ratio_{var}(O_{M_j}, R_{M_j}) \tag{8}$$

Average Time-anchored Rank (ATR)

Since the relevance of meanings is not constant over time, we define the popularity of a meaning in a point of time, $popularity_{M_j}(t)$. A lexical expression can potentially denote multiple meanings, each characterized with a certain degree of time-anchored popularity. Likewise, we order the list of candidate meanings for a given lexical expression based on their popularity at the moment of publishing of the dataset document. For example, if the dataset covers news about a chess tournament, we will see a temporal peak for the chess meaning of *horse* relative to the other meanings. The popularity rank of each meaning, including the correct gold standard meaning, is its position in this ordered list. By averaging over the ranks of all golden candidates we can compute the Average Time-anchored Rank of the golden candidates in a dataset, which gives an indication about the relation between the relative temporal popularity of a meaning and the probability that it is the correct interpretation of an expression, varying from stable to extremely dynamic relations. An ATR rank of a dataset close to 1 indicates a strong bias towards the popular meanings at the time of creation of the dataset.

Average Time-anchored Relative Frequency of Usage (ATRFU)

The potential bias of meaning dominance with respect to its temporal popularity can alternatively be assessed through its frequency of usage at a point of time. We denote the usage of a meaning with U_{M_j}. For a given lexical expression, we compute the relative temporal frequency of usage (FU) of the golden meaning relative to the frequency of usage of all candidate meanings:

$$FU_{M_j}(t) = \frac{U_{M_j}(t)}{\sum\limits_{i=1}^{n} U_{M_i}(t)} \tag{9}$$

The average relative frequency of usage at a given time point (ATRFU) is an average of the frequency values of all gold standard meanings in a dataset. We introduce this metric in order to gain insights into the popularity difference between the competitive meanings at a given time period. This metric would allow us, for instance, to detect that in July 2014 *the United States men's national soccer team* was much more popular than *the women's national soccer team*, while *Tesla Motors* was only slightly more popular than *Nikola Tesla* in May 2015.

Dataset Time Range (DTR)

We define DTR as a time interval between the earliest and the latest published document of a dataset:

$$DTR = [min(date_{doc}), max(date_{doc})] \tag{10}$$

where $date_{doc}$ is the publishing date of a document. For instance, the DTR of the MEANTIME (Minard et al., 2016b) dataset is $[2004, 2011]$.

5.2 Tasks

We demonstrate the applicability of the metrics defined in Section 5.1 on a selection of disambiguation tasks. We cover both concept-oriented tasks (WSD and SRL), as well as instance-based tasks (EL, EnC, and EvC).[3] In Table 1, we specify the model components per disambiguation task, enabling the metrics to be computed. The metrics concerning lexical resources (WordNet (Fellbaum, 1998) for WSD, and PropBank (Kingsbury and Palmer, 2002) for SRL) are only computed for the concept-oriented tasks. Whereas lexical resources, such as WordNet and PropBank, can be seen as reasonable proxies for most of the expressions and concepts known to a generation, it is more difficult to consider databases of instances, such as DBpedia,[4] to approximate all the possible instances that expressions, e.g. *Ronaldo*, can refer to. This is especially the case for events, e.g. the goals *Ronaldo* scored, or the *Ronaldo* t-shirts being sold in a fan shop. There is hardly any registry of real world events independent of the mentions of events in text. Likewise, we only find a few *Ronaldo* entities in DBpedia. Despite its impressive size, DBpedia only covers a very small subset of all instances in the world.

Task	Lexical expression	Meaning	Resource
WSD	lemma	sense	WordNet
SRL	predicate mention	predicate	PropBank
EL	entity mention	entity	DBpedia
EnC	entity mention	entity	DBpedia
EvC	event mention	event	/

Table 1: Task specification of model components.

5.3 Datasets

The choice of datasets conforms to the following rationale. We consider test datasets with running text in English,[5] because we assume that they are the most natural instantiations of the interaction between lexical expressions and meanings and tend to report on the changes in the world. Moreover, such datasets lend themselves better for joint tasks. Finally, we favor publicly available datasets which are commonly used in recent research.

The chosen datasets per disambiguation task are as follows.

WSD The following datasets were taken into consideration: Senseval–2 (**SE2 AW**): All-Words task (Palmer et al., 2001) ; Senseval-3 (**SE3 task 1**): Task 1: The English all-words task (Snyder and Palmer, 2004) ; SemEval-2007 (**SE7 task 17**): Task-17: English Lexical Sample, SRL and All Words (Pradhan et al., 2007) ; SemEval–2010 (**SE10 task 17**): Task 17: All-Words Word Sense Disambiguation on a Specific Domain (Agirre et al., 2010); SemEval–2013 (**SE13 task 12**): Task 12: Multilingual Word Sense Disambiguation (Navigli et al., 2013). The number of test items per competition ranges from roughly 500 to 2500 instances. All most frequent sense baselines are around 65%, except for **SE10 task 17**, in which the focus was on domain-specific WSD, resulting in a most frequent sense baseline of 55%.

SRL For Semantic Role Labelling, we selected the CoNLL-2004 Shared Task: Semantic Role Labeling (**CoNLL04**) (Carreras and Màrquez, 2004). In total, 9,598 arguments were annotated for 855 different verbs.

EL We consider the following datasets: AIDA-YAGO2 (**AIDA test B**) (Hoffart et al., 2011), **WES2015** (Waitelonis et al., 2015), and **MEANTIME** (Minard et al., 2016b). We analyze the commonly used test B collection from the AIDA-YAGO2 dataset, which contains 5,616 entity expressions in 231 documents. WES2015 contains 13,651 expressions in 331 documents about science, while the MEANTIME corpus consists of 120 documents regarding four topics, with 2,750 entity mentions in total.

EnC Guha et al. (2015) created a dataset, QuizBow, for nominal coreference, containing 9,471 mentions in 400 documents. The data annotated comes from a game called quiz bowl.[6]

[3] Note that in the case of SRL we focus on the expression-to-meaning mapping of predicates and do not analyze roles.

[4] `http://dbpedia.org`

[5] Our analysis in this paper is performed on 13 English datasets. The metrics we define in Section 5.1 can easily be applied to many other languages. Namely, the resource-dependent metrics (RORA and RORV) can be applied to the wide range of languages in which DBpedia/WordNet/PropBank are available (for an illustration, DBpedia is currently available in 125 languages). Furthermore, all other metrics rely solely on the annotated textual content within a corpus, which makes them applicable for any language.

[6] `https://en.wikipedia.org/wiki/Quiz_bowl`

EvC we consider three event coreference corpora: EventCorefBank (**ECB**) (Lee et al., 2012), **ECB+** (Cybulska and Vossen, 2014), and EventNuggets (**TAC KBP '15**) (Mitamura et al., 2015). ECB contains 480 documents spread over 43 topics, while its extension ECB+ contains an additional 502 documents spread over the same set of topics. The training corpus of TAC KBP '15 contains 7,478 event coreference chains (hoppers).[7]

6 Analysis

In this Section, we study to what extent datasets cover the complexity of the disambiguation task.[8]

Task	Dataset	MOA	MOV	MODA	MODV	EMNLE	ELENM
WSD	SE2 AW	1.20	1.06	0.94	0.98	0.13	0.05
	SE3 task 1	1.21	1.05	0.94	0.98	0.13	0.04
	SE7 task 17	1.14	1.04	0.95	0.98	0.10	0.03
	SE10 task 17	1.25	1.06	0.93	0.98	0.13	0.05
	SE13 task 12	1.10	1.06	0.97	0.98	0.14	0.05
SRL	CoNLL04	1.20	1.00	0.96	1.00	0.09	0.00
EL	AIDA test B	1.09	1.35	0.98	0.91	0.05	0.22
	WES2015	1.06	1.33	0.97	0.88	0.05	0.21
	MEANTIME	1.19	4.63	0.98	0.64	0.04	0.55
EnC	QuizBowl	1.59	1.80	0.92	0.74	0.13	0.46
EvC	ECB	1.61	3.87	0.89	0.61	0.19	0.65
	ECB+	2.09	3.40	0.85	0.66	0.27	0.57
	TAC KBP '15	4.97	1.22	0.69	0.94	0.47	0.12

Table 2: Observed ambiguity, variance and dominance.

According to Table 2, high complexity in both directions, i.e. high ambiguity and variance, is rare, though the extent of this complexity varies per task. The datasets evaluating WSD, SRL, and EL almost have a 1-to-1 mapping between lexical expressions and meanings, while coreference datasets have higher ambiguity and variance. This can be due to the following reasons: 1. Some of the coreference datasets deliberately focus on increasing ambiguity. 2. An inherent property of coreference seems to be high variance. Similarly, our dominance metrics (MODA/MODV and EMNLE/ELENM) demonstrate a strong bias in our datasets: typically, for any of the datasets, approximately 90% of the occurrences belong to the dominant class on average.

Task	Dataset	ATR	ATRFU
EL	WES2015	1.92	0.53
EL	MEANTIME	1.51	0.51

Table 3: ATR and ATRFU values of the datasets.

Task	Dataset	RORA	RORV
WSD	SE2 AW	0.26	0.38
	SE3 task 1	0.23	0.37
	SE7 task 17	0.20	0.36
	SE10 task 17	0.25	0.40
	SE13 task 12	0.26	0.40
SRL	CoNLL04	0.63	1.00

Table 4: RORA and RORV values of the datasets.

Concerning the concept-oriented tasks, Table 4 shows a notable difference in the complexity of the interaction between the proxies of datasets and resources.[9] Between 74 and 80% of the resource ambiguity per expression is not represented in the datasets, whereas this is the case for 60-64% of the resource

[7] We were unable to obtain the test data for the TAC KBP '15 dataset, hence our analysis is performed on the training data.

[8] The metrics and the analyses of the datasets can be found at `https://github.com/cltl/SemanticOverfitting`.

[9] While computing RORA and RORV, we ignore cases with resource ambiguity and variance of 1.

variance per concept. This is an indication of strong semantic overfitting of the data to a small selection that is not representative for the full potential of expressions and meanings. Furthermore, we observe that this representativeness is relatively constant across concept datasets, which in part can be explained by the fact that the WSD and SRL datasets mainly stem from the same time period (Figure 2), and even from the same corpus (Hovy and Søgaard, 2015). One could argue that the data is correctly representing the natural complexity of a specific time period and genre but it does not challenge systems to be able to shift from one situation to another. We also note a temporal discrepancy between the concept- and instance-based datasets, with the instance-based systems being evaluated on more recent data.

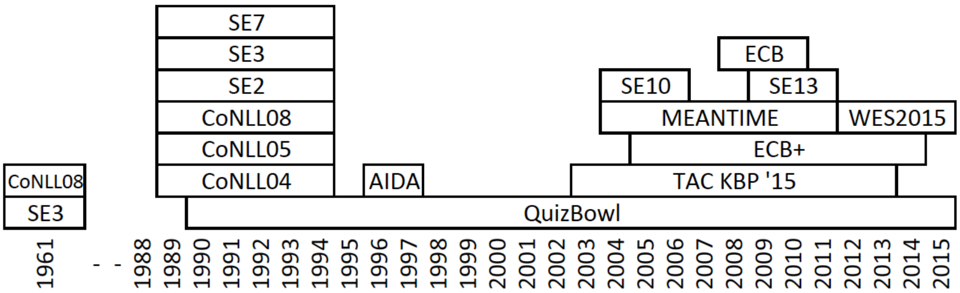


Figure 2: DTR values of the datasets

To understand further the time-bound interaction in our datasets, we study them together with time-bound resources. While our lexical resources and instance knowledge sources contain very little temporal information, we rely on query monitoring websites (Wikiviews and GoogleTrends) to get an indication of the usage of a meaning over time. In Table 3, we show the temporal popularity of entities among their candidates in our datasets according to our web sources.[10] We note a correspondence between the dominance of entities in datasets and their frequency of usage at that time, which exposes a bias of existing datasets towards the most popular entities at the time of their creation.

Our analysis reveals that the existing disambiguation datasets show a notable bias with respect to the aspects of ambiguity, variance, dominance, and time, thus exposing a strong semantic overfitting to a specific part of the world, while largely ignoring long tail phenomena. Typically this is the part of the world that is best known within the context of a generation at a moment of time. This implies that our datasets have a strong bias towards meanings that are popular at that particular moment in time and do not represent the temporal relativity of this bias. Although our metrics provide us with a valuable set of insights into the evaluation datasets, complementary statistical measures should be introduced in the future to capture individual distinctions blurred by averaging over a dataset. These could measure the distribution of ambiguity and variance, their relation to dataset size, and outliers.

7 Proposal for improving evaluation

The direct contribution of our work lies in metric-based evaluation of datasets and resources for systems, which helps interpreting their ability to cope with alterations of ambiguity, variance, dominance, and time. Provided that a collection of multi-task annotated data is available at a central place, our metrics could be applied to output a dataset following certain criteria, e.g. a test set annotated with WSD and EL, whose ambiguity and variance are both between 1.2 and 1.4, and whose documents have been created in the 90s. The practical obstacle is the availability of input data, which can be addressed by the following (semi)automatic expansion method: 1. Collect annotated data and split the data according to time periods. 2. Collect annotated expressions from the data with their dominant meanings. 3. Retrieve

[10] Due to the non-availability of information for the other tasks, we only analyze the temporal dominance for the EL task, even though the set of represented entities in DBpedia is not complete (as discussed in Section 5.3). In our analysis, we only consider ambiguous expressions that can denote more than one entity candidate. The candidates were obtained from the Wikipedia disambiguation pages. From Wikiviews, the month of the document creation time was used for the dominance information.

new documents using unsupervised techniques in which these expressions occur with evidence for usage in other meanings than the dominant one in the existing datasets. Evidence can come from meta data, unsupervised clustering, and temporal and topical distance from annotated data. 4. Fix alternative meanings for all tokens in the new texts (one meaning-per-document), if necessary applying additional disambiguation tools. Add this data as silver data to the collection. 5. If necessary, re-annotate silver data manually or add annotations for other tasks.[11] 6. Spread documents over different time periods for both annotated gold data and silver data to obtain sufficient variation in time-bound contexts. Provided that this acquisition procedure is successful, selecting a dataset would require almost no effort, which enables creation of many, well-motivated datasets. Consequently, the dynamic nature of this process would challenge semantic overfitting.

In case the proposed data acquisition procedure proves too hard or too laborious to realize, we propose an alternative, namely an event-based Question Answering (QA) task, extensively elaborated on in Postma et al. (2016), whose dataset would contain documents gathered in a smart automatic way. The data acquisition procedure for this dataset is driven by multiple confusion factors: ambiguity, dominance, variance, time, location, and topic. This data would reflect a high degree of ambiguity and variance and would capture a wide range of small real-world phenomena. In order to perform on this task with a good accuracy, the systems will be required to exhibit a deeper semantic understanding of the linguistic tail of the disambiguation tasks we analyze in this paper. However, the only task that will explicitly be evaluated is the QA task itself, which means that the annotation task would be largely reduced to the components necessary for the questions and answers.

The resulting time-aware evaluation datasets, originating from both the annotation-based and QA-based approaches, allow the community to test understanding of language originating from different generations and communities, and a community's language usage in relation to different world contexts. It would also assess to what extent a disambiguation system can adapt to language use from another time slice than the one trained on, with potentially new meanings and expressions, and certainly a different distribution of the expression-meaning relation. We believe this challenges semantic overfitting to one single part and time of the world, and will inspire systems to be more robust towards aspects of ambiguity, variance, and dominance, as well as their temporal dependency.

8 Conclusion

We qualified and quantified the relation between expressions and meanings in the world for a generation sharing a language system, as well as the fluctuation in usage of expressions and meanings over time. We proposed the Semiotic Generation and Context Model, which captures the distribution changes in the semiotic relation given the context of the changing world. We apply it to address three key problems concerning semantic overfitting of datasets. We conceptualize and formalize generic metrics which evaluate aspects of datasets and provide evidence for their applicability on popular datasets with running text from five disambiguation tasks. We observe that existing disambiguation 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. Finally, we propose a time-based, metric-aware approach to create datasets in a systematic and semi-automated way as well as an event-based QA task. Both approaches will result in datasets that would challenge semantic overfitting of disambiguation systems.

References

Eneko Agirre, Oier López de Lacalle, Christiane Fellbaum, Shu-Kai Hsieh, Maurizio Tesconi, Monica Monachini, Piek Vossen, and Roxanne Segers. 2010. SemEval-2010 Task 17: All-Words Word Sense Disambiguation on a Specific Domain. In *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval-2010)*, pages 75–80, Uppsala, Sweden, July. Association for Computational Linguistics.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp

[11] Excessive labor could be avoided by prioritizing relevant expressions, e.g. according to the metrics.

Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop & Interoperability with Discourse*, pages 178–186.

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.

Xavier Carreras and Lluís Màrquez, 2004. *Proceedings of the Eighth Conference on Computational Natural Language Learning (CoNLL-2004) at HLT-NAACL 2004*, chapter Introduction to the CoNLL-2004 Shared Task: Semantic Role Labeling.

Wanxiang Che and Ting Liu. 2010. Jointly modeling WSD and SRL with Markov logic. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, pages 161–169. Association for Computational Linguistics.

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.

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.

Christiane Fellbaum. 1998. Wordnet: an electronic lexical database. *MIT Press, Cambridge MA*, 1:998.

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)*.

Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordin, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. Robust Disambiguation of Named Entities. In *Conference on Empirical Methods in Natural Language Processing*, EMNLP.

Dirk Hovy and Anders Søgaard. 2015. Tagging performance correlates with author age. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 483–488.

Ioana Hulpuş, Narumol Prangnawarat, and Conor Hayes. 2015. Path-based semantic relatedness on linked data and its use to word and entity disambiguation. In *Proceedings of the International Semantic Web Conference (ISWC)*, pages 442–457. Springer.

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.

Paul Kingsbury and Martha Palmer. 2002. From Treebank to Propbank. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*.

Heeyoung Lee, Marta Recasens, Angel Chang, Mihai Surdeanu, and Dan Jurafsky. 2012. Joint entity and event coreference resolution across documents. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2012)*, pages 489–500. Association for Computational Linguistics.

Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll. 2004. Finding predominant word senses in untagged text. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics (ACL 2004)*, page 279. Association for Computational Linguistics.

Anne-Lyse Minard, Manuela Speranza, Ruben Urizar, Begona Altuna, Marieke van Erp, Anneleen Schoen, and Chantal van Son. 2016a. MEANTIME, the newsreader multilingual event and time corpus. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*.

Anne-Lyse Minard, Manuela Speranza, Ruben Urizar, Begona Altuna, Marieke van Erp, Anneleen Schoen, and Chantal van Son. 2016b. Meantime, the newsreader multilingual event and time corpus. *Proceedings of LREC2016*.

Teruko Mitamura, Zhengzhong Liu, and Eduard Hovy. 2015. Overview of tac-kbp 2015 event nugget track. In *Text Analysis Conference*.

Andrea Moro and Roberto Navigli. 2015. Semeval-2015 task 13: Multilingual all-words sense disambiguation and entity linking. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval-2015)*, pages 288–297.

Andrea Moro, Alessandro Raganato, and Roberto Navigli. 2014. Entity linking meets word sense disambiguation: a unified approach. *Transactions of the Association for Computational Linguistics (TACL)*, 2:231–244.

Roberto Navigli, David Jurgens, and Daniele Vannella. 2013. SemEval-2013 Task 12: Multilingual Word Sense Disambiguation. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 222–231, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

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

Martha Palmer, Christiane Fellbaum, Scott Cotton, Lauren Delfs, and Hoa Trang Dang. 2001. English Tasks: All-Words and Verb Lexical Sample. In *Proceedings of the Second International Workshop on Evaluating Word Sense Disambiguation Systems (SENSEVAL-2)*, pages 21–24, Toulouse, France, July. Association for Computational Linguistics.

Marten Postma, Filip Ilievski, Piek Vossen, and Marieke van Erp. 2016. Moving away from semantic overfitting in disambiguation datasets. In *Proceedings of the EMNLP Workshop on Uphill Battles in Language Processing*.

Sameer Pradhan, Edward Loper, Dmitriy Dligach, and Martha Palmer. 2007. SemEval-2007 Task-17: English Lexical Sample, SRL and All Words. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 87–92, Prague, Czech Republic, June. Association for Computational Linguistics.

Sameer Pradhan, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue. 2011. CoNLL-2011 Shared Task: Modeling unrestricted coreference in OntoNotes. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning (CoNLL 20111)-Shared Task*, pages 1–27. Association for Computational Linguistics.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 Shared Task: Modeling multilingual unrestricted coreference in OntoNotes. In *Joint Conference on Empirical Methods on Natural Language Processing (EMNLP) and on Natural Language Learning (EMNLP-CoNLL 2012)-Shared Task*, pages 1–40. Association for Computational Linguistics.

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.

Benjamin Snyder and Martha Palmer. 2004. The English All-Words Task. In Rada Mihalcea and Phil Edmonds, editors, *SensEval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text.*, pages 41–43, Barcelona, Spain, July. Association for Computational Linguistics.

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)*.

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.

Jörg Waitelonis, Claudia Exeler, and Harald Sack. 2015. Linked Data Enabled Generalized Vector Space Model to Improve Document Retrieval. In *Proceedings of NLP & DBpedia 2015 workshop in conjunction with 14th International Semantic Web Conference (ISWC2015), CEUR Workshop Proceedings*.