DEFT: A corpus for definition extraction in free- and semi-structured text

Sasha Spala[1], **Nicholas A Miller**[1], **Yiming Yang**[2*], **Franck Dernoncourt**[3], **Carl Dockhorn**[1]

[1]Adobe Inc.
345 Park Ave.
San Jose, CA
{sspala,
nimiller,
cdockhorn}
@adobe.com

[2]University of California, San Diego
9500 Gilman Dr.
La Jolla, CA
yiy001
@eng.ucsd.edu

[3]Adobe Research
345 Park Ave.
San Jose, CA
dernonco
@adobe.com

Abstract

Definition extraction has been a popular topic in NLP research for well more than a decade, but has been historically limited to well-defined, structured, and narrow conditions. In reality, natural language is messy, and messy data requires both complex solutions and data that reflects that reality. In this paper, we present a robust English corpus and annotation schema that allows us to explore the less straightforward examples of term-definition structures in free and semi-structured text.

1 Introduction

As the computational linguistics community moves further towards comprehensive natural language understanding, it has become increasingly clear that our methods need to consider scenarios that match a complex linguistic reality. In the case of term-definition pairs, that means exploring how explicit in-text definitions and glosses work in free and semi-structured text, especially those whose term-definition pair span crosses a sentence boundary and those lacking explicit definition phrases. In this paper we present a new corpus of natural language term-definition pairs, as well as a novel schema that can be generally applied for a wide range of domains.

2 Related Work

Most related work on definition extraction has relied on the idea that definitions can be captured by common "definitor" verb phrases like "means", "refers to", and "is". Early work in the field incorporated rule-based methods that extracted sentences that met this narrow standard (JL Clavens, 2001; Cui and Chua, 2004, 2005; Fahmi and Bouma, 2006; Zhang and Jiang, 2009). While predictable and easily applied, these models subsequently failed to extract sentences that lack these explicit markers. In an effort to expand on the type of phrases used to extract definitions, Cui et al. (2007) used soft pattern matching in a modified HMM (PHMM). More recent work from Espinosa Anke and Schockaert (2018) makes use of a neural approach, which reached state-of-the-art performance on the word class lattices (WCL) datasets (Navigli et al., 2010). Even so, these methods require both term and definition to appear in the same sentence and for terms to appear before definitions.

Hypernym detection, a related field, has also garnered interest for quite some time (see e.g., Hearst (1992); Snow et al. (2005); Ritter et al. (2009); Shwartz et al. (2017)). Because many hypernym glosses follow the pattern *X, such as Y* or *X is a (type of) Y*, this work contains a subset of cases considered for definition extraction. Navigli and Velardi (2010) demonstrated the use of word class lattices for *both* hypernym detection and definition extraction, and Yin and Roth (2018) proved the effectiveness of including definitions in the training of hypernym detection models.

Most work on definition extraction has been applied solely to English datasets, including the WCL dataset mentioned above (Navigli et al., 2010), the ukWaC dataset (Ferraresi et al., 2008), a large crawled dataset of the .uk domain name, and the W00 dataset, a small, expertly annotated corpus introduced by Jin et al. (2013). There does exist a smaller effort for multilingual explorations, including German (Storrer and Wellinghoff, 2006), Portuguese (Del Gaudio and Branco, 2007), and Slavic (Przepiórkowski et al., 2007), as well as some language-independent approaches (Del Gaudio and Branco, 2009). The vast majority of these approaches are for unstructured text, typically scraped from online sources, as in the ukWaC dataset, though some interest has been given specifically for semi-structured text in legal contracts (see e.g. Curtotti and McCreath

*Work was completed while individual was employed at Adobe Research.

Proceedings of the 13th Linguistic Annotation Workshop, pages 124–131
Florence, Italy, August 1, 2019.

Dataset	# of positive annotations	Size (in sentences)
WCL	1,871	4,718
W00	731	2,185
DEFT	**11,004**	**23,746**

Table 1: Definition extraction datasets

(2010) and Winkels and Hoekstra (2012)).

While variations of the *X is a Y* form are indeed common definition sentence structures, they do not capture a wide range of definition structures that appear in both free and semi-structured text. In particular, they typically constrain the environment in which we find these definitions. We see this in cases like the WCL dataset, of which a portion of the data was extracted by taking the first sentences of randomly sampled Wikipedia articles, as well as in much of the legal domain research, which often consider only the definitions which appear in explicitly-identified glossary sections'. Our proposed Definition Extraction from Texts (DEFT) corpus aims to alleviate this problem by providing complex, human-annotated data across a variety of topics and among both free (textbook) and semi-structured (legal document) language.

3 Corpus

The DEFT corpus[1] consists of annotated content from two different data sources: 1) 2,443 sentences (5,324,430 tokens) from various 2017 SEC contract filings from the publicly available US Securities and Exchange Commission EDGAR (SEC) database, and 2) 21,303 sentences (409,253 tokens) from the https://cnx.org/ open source textbooks (by various authors, licensed under CC BY 4.0) including topics in biology, history, physics, psychology, economics, sociology, and government. 22% of SEC sentences contain definitions and 28% of textbook sentences contain definitions. Our entire corpus, including both datasets, is significantly larger and more complex than any existing definition extraction dataset (see Table 1).

During annotation, we found that roughly 50% of term-definition pairs appeared across sentence boundaries or with an otherwise complex structure (e.g., containing secondary information, containing ambiguous references to previously stated terms or definitions) whereby the relationship between a term and definition requires more deduction than finding a definition verb phrase.

Our annotation schema is outlined in Table 2 and Table 3. Terms, alias terms, referential terms, and ordered terms are always annotated as a complete NP, including any determiner that may appear with the noun. Where possible, definitions, secondary definitions, referential definitions, and ordered definitions consume the entire clause(s) in which they appear. Qualifiers, which were added to handle date, location, and condition nuances in legal language, are also annotated at the clause level. Terms may not exist without either a matching alias term or definition.

With the exception of the qualifier tag, which appears only in the SEC data, the schema is applied generally across both datasets.

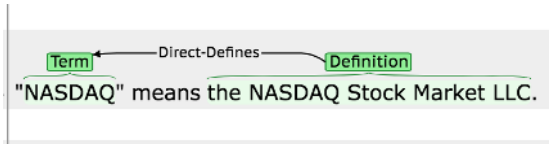

Figure 1: A typical definition within the "Definitions" section of a legal contract.

3.1 Annotation Schema

As mentioned above (see Section 2), previous work has focused primarily on term-definition pairs that appear in the same sentence. Navigli et al. (2010) used a formalized schema from Storrer and Wellinghoff (2006), which identifies a *definiendum*, *definitor*, *definiens*, and *rest* field for each term-definition pair. Curtotti and McCreath (2010) use "definition clauses", drawing on definitions in a legal sense - that is, those which appear in a formal definition or glossary section and which do not cross sentence boundaries. These definition clauses typically encompass an entire sentence; the matching term either appears in context (within the natural language of the definition clause) or with some formatting (e.g. bold, italic, heading-like) to indicate its relationship with the definition clause.

Our schema expands on these strategies to account for a wider variety of term-definition structures. Because of the sweeping variety of "definition-like" verb phrases (e.g. *means, is, defines*, etc.), and the apparent lack thereof in some

[1] https://github.com/adobe-research/deft_corpus

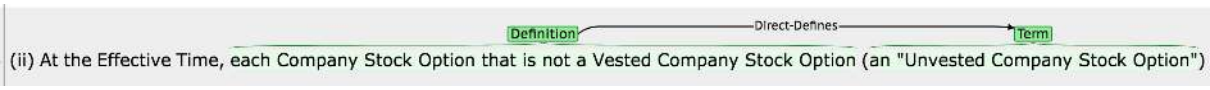

Figure 2: A typical "gloss" in the body of a contract, where a term is identified by enclosing parentheses and quotations which separate it from its definition.

Figure 3: A typical preamble of a contract in the SEC filings, with qualifiers that clarify the date, location, or condition in which the term appears.

cases (see e.g. Figs 2, 3, 7), we are most interested in identifying terms and definitions, but not necessarily the verb phrases which may or may not connect the two.

Annotators were instructed to identify definitions that had an explicitly mentioned referring term. Definitions may span entire sentences, or may be a single clause-level or smaller NP. In our schema, definitions are not merely general descriptions of a term, but refer clearly back to the term they define, and can define *only* the term. If it does not already appear as such, the term and chosen definition sequence can be typically rephrased as *X is a Y*. Definitions do *not* include definitors, words that introduce restrictive or non-restrictive clauses (such as *that, who, which*), or narratives. Definitions must also be apparent from the explicitly written text available to the annotator. Our guidelines avoid "implicit" definitions, or definitions that require external understanding of the topic to parse. If a definition crosses a sentence boundary, the sequence (in some cases, a full sentence) following the boundary identified as definition-like is labelled as a secondary definition.

3.2 Contract Data

As mentioned above, the corpus consists of 2443 sentences from SEC contract filings. These sentences are often long, with several term-definition pairs appearing within one sentence. While it is well known that many contracts contain "definition sections", glossaries, or definition clauses (Curtotti and McCreath, 2010; Curtotti and Sridharan, 2013), our annotation efforts revealed that in reality, definitions appear throughout the entire contract. Because of the nature of this spread, our annotators were instructed to annotate entire legal documents, not just the labeled "definition" sections as in Fig 1. Often, glosses outside definition sections are identified by a term that appears in quotations and bounded by parentheses, separating it from the inline text (see Fig 2). Occasionally, the inline definitions use referential terms or definitions to indirectly define primary terms, though this is a rare case ($< 1\%$ of tags).

As mentioned above, the SEC data includes the qualifier tag, which is often found qualifying terms or alias terms in contract preambles, as seen in Fig 3. These preambles commonly contain terms with matching alias terms, but no explicit definition. It is also important to note that terms in these preambles are typically *not* the longer, expanded acronym, or more formal representation (as they may be in the textbook data, by nature of how the textbooks' style refers to those terms), but rather the acronym or otherwise shortened form of the term, as this is how they are referred to throughout the rest of the document. Here we see an interesting divergence between the two domains: In textbooks the goal may be to educate the reader of the term, and thus often uses the more formalized representation, but in contracts the goal is usually clarity, brevity, and adherence to legal code.

3.3 Textbook Data

In the textbook data, three-sentence context windows were sampled from sentences that contained a bold n-gram (a strong signal in educational texts indicating a formally defined term) with a context sentence on either side of the sentence with those bold token(s). Consistent with previous research (Cui et al., 2007; Degorski and Przepiorkowski, 2008; Curtotti and McCreath, 2010; Navigli et al., 2010) definitions do in fact appear in the *X is a Y* form, with a clear "definitor". However, many textbook examples also lack this explicit trigger, and instead implicitly define the relationship between the term and definition, either by a referential term or referential definition, or through the

Tag Name	Description
Term	A primary term
Alias Term	A secondary, less common name for the primary term. Links to a term tag.
Ordered Term	Multiple terms that have matching sets of definitions which cannot be separated from each other without creating an non-contiguous sequence of tokens. E.g. *x and y represent positive and negative versions of the definition, respectively*
Referential Term	An NP reference to a previously mentioned term tag. Typically *this/that/these + NP*
Definition	A primary definition of a term. May not exist without a matching term.
Secondary Definition	Supplemental information that may qualify as a definition sentence or phrase, but crosses a sentence boundary.
Ordered Definition	Multiple definitions that have matching sets of terms which cannot be separated from each other. See Ordered Term.
Referential Definition	NP reference to a previously mentioned definition tag. See Referential Term.
Qualifier	A specific date, location, or condition under which the definition holds

Table 2: Tag schema

Relation Name	Description
Direct-defines	Links definition to term.
Indirect-defines	Links definition to referential term or term to referential definition.
Refers-to	Links referential term to term or referential definition to definition.
AKA	Links alias term to term.
Supplements	Links secondary definition to definition.

Table 3: Relation schema

2267.
So too did the appointment of Clay as secretary of state.
<START> John C. Calhoun labeled the whole affair a "corrupt bargain" ([link]).<END>.
Everywhere, Jackson supporters vowed revenge against the anti-majoritarian result of 1824.

Figure 4: An excerpt from the extracted textbook sentences without a term-definition pair.

implication of the syntactic structure (see e.g., Fig 7). It is important to note that, as seen in Fig 6, the *X is a Y* form (or some variant thereof) may still appear between the referential term or definition and the primary term or definition, especially when the relationship between the primary term and primary definition crosses a sentence boundary.

Though we may not have captured all examples of term-definition pairs in textbooks, this did allow us to regularly and implicitly, without active annotator tagging, identify examples which may appear to be definitions at a surface level, but in fact, do not meet our schema criteria for a definition. In particular, because of the constraints of our schema and the unclear ground truth definition of people and places, our annotation excludes these cases. With the exception of definitions including the formal title of an individual or the physical composition of a location (see, e.g. Fig 4, Fig 5), they are not included in the corpus. All three-sentence windows that appear in the dataset without any labels are considered false positives, as they do contain bold tokens, but either do not have distinguishable definitions or provide auxiliary information not integral to the ground truth definition of the term, as in the case of people and places.

4 Annotation Process

The data in this corpus was annotated by a total of five annotators using the brat annotation framework (Pontus Stenetorp and Tsujii, 2012). A group of three annotators labeled data from the textbook corpus and another group of three annotators labeled the contract data, with one anno-

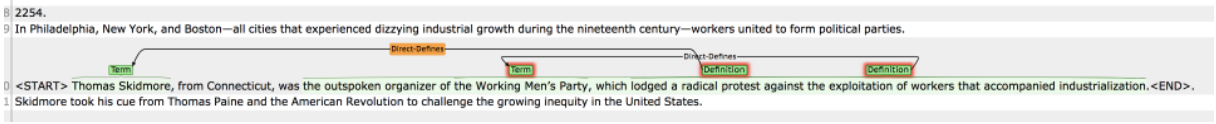

Figure 5: A person labeled as a term with a qualifying definition.

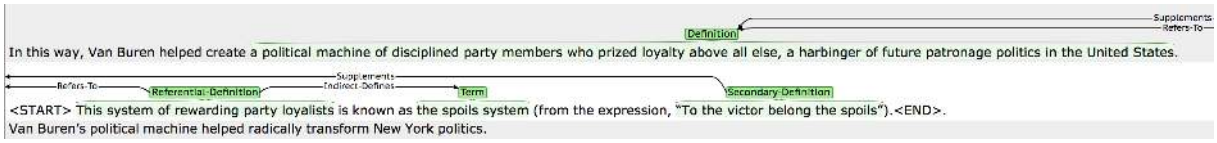

Figure 6: A cross-sentence term-definition pair, where the definition appears before the statement of the term and additional definition information is provided in the form of a secondary definition.

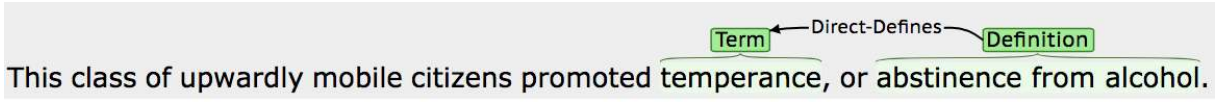

Figure 7: A term-definition pair where the term is implied to be related to the definition by way of clausal separation.

tator having also labeled the textbook data. The development of the annotation schema followed the MAMA cycle (Pustejovsky and Stubbs, 2013), with an emphasis on providing the most pragmatic annotation process while still capturing the most accurate representations of generalized definition structures. Annotators were trained before beginning annotation on the textbook data, then again before beginning annotation on the contract data.

4.1 Inter-annotator Agreement

Inter-annotator agreement (IAA) is measured using a modified version of Krippendorff's alpha (Krippendorff, 2011) with the MASI distance metric (Passonneau, 2006) in order to account for and score partial sequential overlaps of text:

$$\delta(c,k) = \begin{cases} MASI(c,k), & \text{if } c = k \\ 1, & \text{otherwise} \end{cases}$$

Where $c = k$ and the text spans match exactly, the MASI distance is 0.

IAA was calculated after every training period, with a final annotator agreement score of $\alpha_{term} = 0.80$ and $\alpha_{definition} = 0.50$ for the textbook corpus and $\alpha_{term} = 0.85$ and $\alpha_{definition} = 0.54$ for the contract corpus. We believe these IAA scores match the reality of human performance on such a complicated task. After training time, each sentence in the corpus was labeled by one annotator. For the textbook annotation, each annotator was assigned a list of three-sentence passages randomly distributed from every textbook topic. For the contract annotation, each annotator was assigned a set of a set of whole contracts to annotate.

4.2 Annotation Challenges

Though our annotation schema is intended to apply to cross-domain definition extraction, there are still certain linguistic differences between the two data sources. In particular, the goals of different document types and formats seems to instruct the use of definitions in their contexts. We briefly discussed a symptom of this in section 3.2, where the primary term takes different levels of formality depending on the intent of the document: in contracts, it is typically the simplified, abbreviated form, and in textbooks it is typically the expanded or formalized representation. We believe the same influence drives the appearance of the qualifier construct in contracts. Legal documents, by necessity, must state the conditions under which a trait, event, or system is true. This often presents as a relevant date (before, after, or on which the terms apply) or location (such as a country or state under which the terms apply). Textbooks, on the other hand, do not require this level of specificity; though they may state similar facts, such as the date or location of an event, this information is arguably not crucial to the understanding of the core definition. While we may argue for either including or excluding these textbook counterparts, the DEFT corpus does not label them. Our annotation process favors maintaining the most basic definition of the term without compromising

Figure 8: An example where the definition is implied from the legal "force" of the contract.

necessary information. Legal contracts also occasionally "define" terms implicitly by the legal "force" of the document. In Fig 8, "Company's Default" is an event that happens when the company's Loan Documents default. However, this does not directly define what a default is, only the implied conditions under which it happens. From the formatting of the sentence, it is clear that the author intends for "Company's Default" to be a term. "Company's Default" is indeed referred to by name later in the same document. However, it is assumed that the reader has enough knowledge of the process of defaulting that they may infer what the Company's Default means in this context.

Textbooks have similarly difficult terms: people and places, briefly discussed in section 3.3. These terms appear bold, implying the same author intent as the parenthesized terms in the legal contracts. However, the definition of these terms remain vague: is a person defined by their most well-known achievement (especially in historical contexts)? Are they defined by where they were born or died? Are places defined by their most common use? Perhaps their location within a larger geographical structure? In many cases, the way in which these examples are "defined" in the text depends on the context in which they are presented; A history textbook detailing the contributions of a major political figure may "define" that individual by their successes or failures, depending on the perspective of the textbook or the context of the broader section of the document that particular example appears in. Again, this reflects the intent of the document or section as a whole. With the exception of an individual's title and the physical composition of a location (especially in a scientific context), we determine these cases to be out of scope for our current research. As mentioned in Section 3.1, definitions must be able to refer to the term *only*, meaning that most general descriptions of locations or individuals do not qualify as definitions under our schema with one exception: A specific physical descriptor of a location, or a statement of an individual's title. These specific examples both qualify as definitions as they do *not* require external knowledge of the term or concept, and can be directly connected back to their respective terms.

5 Conclusion

We believe that the DEFT corpus, as the largest existing corpus with the express purpose of definition extraction in a wide range of contexts, will be a major contribution to the field. In the process of creating and revising the annotation schema, we have unpacked significant nuances in the linguistic structures and requirements of definitions in a variety of contexts. As a significant increase in size and granularity of past definition extraction corpora, the DEFT corpus will be particularly useful from both corpus linguistics and computational linguistics perspectives. We believe that in addition to the existing annotated textbook and contract data, our schema could be applied to other forms of un- and semi-structured documents. The DEFT corpus and its annotation schema are an expansion on the existing assumptions of simple, hypernym-like, definition syntax, and offer a new perspective for the next generation of definition extraction models.

6 Acknowledgements and Licensing

We would like to acknowledge the contributions of our annotation team, Lucino Chiafullo, Micaela Kaplan, Roger LaCroix, Molly Moran and Jennifer Pei-Hsuan Lee, without which we would not have the annotated data we have today.

References

Hang Cui, Min-Yen Kan, and Tat-Seng Chua. 2007. Soft pattern matching models for definitional question answering. *ACM Trans. Inf. Syst.*, 25(2).

Kan M.Y. Cui, H. and T.S. Chua. 2004. Unsupervised learning of soft patterns for generating definitions from online news. In *Proceedings of WWW*, 90-99.

Kan M.Y. Cui, H. and T.S. Chua. 2005. Generic soft pattern models for definitional question answering. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

McCreath Eric Curtotti, Michael and Srinivas Sridharan. 2013. Software tools for the visualization of definition networks in legal contracts. In *Proceedings of the 14th International Conference on Artificial Intelligence and the Law*.

Michael Curtotti and Eric McCreath. 2010. Corpus based classification of text in australian contracts. In *Proceedings of the Australasian Language Technology Association Workshop 2010*.

Micha Marcinczuk Degorski, ukasz and Adam Przepiorkowski. 2008. Definition extraciton using a sequential combination of baseline grammars and machine learning classifiers. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*.

Rosa Del Gaudio and Antonio Branco. 2007. Automatic extraction of definitions in portuguese: A rule-based approach. In *Proceedings of the TeMa Workshop*.

Rosa Del Gaudio and António Branco. 2009. Language independent system for definition extraction: First results using learning algorithms. In *Proceedings of the 1st Workshop on Definition Extraction*, WDE '09, pages 33–39, Stroudsburg, PA, USA. Association for Computational Linguistics.

Luis Espinosa Anke and Steven Schockaert. 2018. Syntactically aware neural architectures for definition extraction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 378–385. Association for Computational Linguistics.

Ismail Fahmi and Gosse Bouma. 2006. Learning to identify definitions using syntactic features. In *Proceedings of the EACL 2006 workshop on Learning Structured Information in Natural Language Applications*.

Adriano Ferraresi, Eros Zanchetta, Silvia Bernardini, and Marco Baroni. 2008. Introducing and evaluating ukwac, a very large web-derived corpus of english.

Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th Conference on Computational Linguistics - Volume 2*, COLING '92, pages 539–545, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yiping Jin, Min-Yen Kan, Jun-Ping Ng, and Xiangnan He. 2013. Mining scientific terms and their definitions: A study of the acl anthology. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 780–790. Association for Computational Linguistics.

S Muresan JL Clavens. 2001. Evaluation of the definder system for fully automatic glossary construction. In *Proceedings of the AMIA Symposium*, pages 324–328.

Klaus Krippendorff. 2011. Computing Krippendorff's alpha-reliability.

Roberto Navigli and Paola Velardi. 2010. Learning word-class lattices for definition and hypernym extraction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 1318–1327, Stroudsburg, PA, USA. Association for Computational Linguistics.

Roberto Navigli, Paola Velardi, and Juana Ruiz-Martnez. 2010. An annotated dataset for extracting definitions and hypernyms from the web. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010*.

Rebecca Passonneau. 2006. Measuring agreement on set-valued items (masi) for semantic and pragmatic annotation.

Goran Topi Tomoko Ohta Sophia Ananiadou Pontus Stenetorp, Sampo Pyysalo and Jun'ichi Tsujii. 2012. brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations Session at EACL 2012*.

Adam Przepiórkowski, Degórski, Beata Wójtowicz, Miroslav Spousta, Vladislav Kuboň, Kiril Simov, Petya Osenova, and Lothar Lemnitzer. 2007. Towards the automatic extraction of definitions in slavic. In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing: Information Extraction and Enabling Technologies*, ACL '07, pages 43–50, Stroudsburg, PA, USA. Association for Computational Linguistics.

James Pustejovsky and Amber Stubbs. 2013. *Natural Language Annotation for Machine Learning*. O'Reilly Media, Inc.

Alan Ritter, Stephen Soderland, and Oren Etzioni. 2009. What is this, anyway: Automatic hypernym discovery. pages 88–93.

Vered Shwartz, Enrico Santus, and Dominik Schlechtweg. 2017. Hypernyms under siege: Linguistically-motivated artillery for hypernymy detection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 65–75, Valencia, Spain. Association for Computational Linguistics.

Rion Snow, Daniel Jurafsky, and Andrew Y. Ng. 2005. Learning syntactic patterns for automatic hypernym discovery. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 1297–1304. MIT Press.

Angelika Storrer and Sandra Wellinghoff. 2006. Automated detection and annotation of term definitions in german text corpora. In *Proceedings of the Conference on Language Resources and Evaluation (LREC)*.

R. Winkels and R. Hoekstra. 2012. Automatic extraction of legal concepts and definitions. In *Legal Knowledge and Information Systems: JURIX 2012: the twenty-fifth annual conference*.

Wenpeng Yin and Dan Roth. 2018. Term definitions help hypernymy detection. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 203–213, New Orleans, Louisiana. Association for Computational Linguistics.

ChunXia Zhang and Peng Jiang. 2009. Automatic extraction of definitions. In *Proceedings of the 2nd IEEE International Conference on Computer Science and Information Technology*.