Weakly Supervised Definition Extraction*

Luis Espinosa-Anke, Francesco Ronzano and **Horacio Saggion**
TALN - DTIC
Universitat Pompeu Fabra
Carrer Tànger, 122-134
08018 Barcelona
{luis.espinosa,francesco.ronzano,horacio.saggion}@upf.edu

Abstract

Definition Extraction (DE) is the task to extract textual definitions from naturally occurring text. It is gaining popularity as a prior step for constructing taxonomies, ontologies, automatic glossaries or dictionary entries. These fields of application motivate greater interest in well-formed encyclopedic text from which to extract definitions, and therefore DE for academic or lay discourse has received less attention. In this paper we propose a weakly supervised bootstrapping approach for identifying textual definitions with higher linguistic variability than the classic encyclopedic *genus-et-differentia* definition, and take the domain of Natural Language Processing as a use case. We also introduce a novel set of features for DE and explore their relevance. Evaluation is carried out on two datasets that reflect opposed ways of expressing definitional knowledge.

1 Introduction

Definition Extraction (DE) is the task to automatically extract textual definitions from text (Navigli and Velardi, 2010). It has received notorious attention for its potential application to glossary generation (Muresan and Klavans, 2002; Park et al., 2002), terminological databases (Nakamura and Nagao, 1988), question answering systems (Saggion and Gaizauskas, 2004; Cui et al., 2005), for supporting terminological applications (Meyer, 2001; Sierra et al., 2006), e-learning (Westerhout and Monachesi, 2007), and more recently for multilingual paraphrase extraction (Yan et al., 2013), ontology learning (Velardi et al., 2013) or hypernym discovery (Flati et al., 2014).

The corpora that have been used for evaluating DE systems are varied, although in general efforts have been greatly focused on academic and encyclopedic genres. Some prominent examples include German technical texts (Storrer and Wellinghoff, 2006), the IULA Technical Corpus (in Spanish) (Alarcón et al., 2009), the ACL Anthology (Jin et al., 2013; Reiplinger et al., 2012), the BNC corpus (Rodríguez, 2004), Wikipedia (Navigli and Velardi, 2010), ensembles of domain glossaries and Web documents (Velardi et al., 2008), or technical texts in various languages (Westerhout and Monachesi, 2007; Przepiórkowski et al., 2007; Borg et al., 2009; Degórski et al., 2008; Del Gaudio et al., 2013).

We propose a DE approach which, from a starting set of encyclopedic definition seeds, self-trains iteratively and gradually fits its classification capability to a target domain-specific test set. Evaluation is carried out on two corpora: First, a set of 50 abstracts of papers in the field of NLP[1]. Here, the target term is defined in the first sentence, and additional information may appear in the form of "syntactically plausible false definitions", i.e. sentences where the target term is also present, relevant information is provided, but do not constitute a definition

* This work is partially funded by the SKATER project, TIN2012-38584-C06-03, Ministerio de Economía y Competitividad, Secretaría de Estado de Investigación, Desarrollo e Innovación, España; and Dr. Inventor (FP7-ICT-2013.8.1 611383).

[1] Henceforth, we refer to this corpus as the *MSR-NLP* dataset.

Proceedings of Recent Advances in Natural Language Processing, pages 176–185,
Hissar, Bulgaria, Sep 7–9 2015.

(Navigli and Velardi, 2010). Second, the *W00* corpus (Jin et al., 2013), a subset of the ACL Anthology manually annotated with definitions, and which includes highly variable definitions both in terms of content and syntax. We achieve competitive results in both corpora.

The main contributions of our paper are: (1) A set of experiments demonstrating the soundness of our approach for DE in two different linguistic registers; (2) A novel set of features and an exploration of their influence in the learning process; and (3) A small, focused benchmarking dataset for DE evaluation in the NLP domain.

The remainder of this paper is structured as follows: Section 2 reviews prominent work in DE; Section 3 provides a detailed description of the datasets used; Section 4 presents the features used in our classification procedure and describes the bootstrapping algorithm; Section 5 shows the performance of our approach; Section 6 lists the best features at important iterations and discusses these findings; and finally Section 7 summarizes the main ideas contained in this paper and outlines potential directions for future work.

2 Background

Definitions are a well-studied topic, which traces back to the Aristotelian genus et differentia model of a definition, where the defined term (*definiendum*) is described by mentioning its immediate superordinate, usually a hypernym (*genus*), and the cluster of words that differentiate such definiendum from others of its class (*definiens*). Furthermore, additional research has elaborated on different criteria to take into consideration when deciding *what is a definition*: either by looking at their degree of formality (Trimble, 1985), the extent to which they are specific to an instance of an object or to the object itself (Seppälä, 2009), the semantic relations holding between definiendum and concepts included in the definiens (Alarcón et al., 2009; Schumann, 2011), the fitness of a definition for target users (Bergenholtz and Tarp, 2003; Fuertes-Olivera, 2010) or their stylistic and domain features (Velardi et al., 2008). In this work we elaborate on some ideas from the latter, especially on their *domain* and *stylistic* filters, which motivated the design of statistically-motivated features to describe a word's salience in terms of definitional knowledge (cf. Section 4).

Regarding DE, the earliest attempts focused on lexico-syntactic pattern-matching, either by looking at cue verbs (Rebeyrolle and Tanguy, 2000; Saggion and Gaizauskas, 2004; Sarmento et al., 2006; Storrer and Wellinghoff, 2006), or other features like punctuation or layout (Muresan and Klavans, 2002; Malaisé et al., 2004; Sánchez and Márquez, 2005; Przepiórkowski et al., 2007; Monachesi and Westerhout, 2008). As for supervised settings, let us refer to (Navigli and Velardi, 2010), who propose a generalization of word lattices for identifying definitional components and ultimately identifying definitional text fragments. Finally, more complex morphosyntactic patterns were used by (Boella et al., 2014), who model single tokens as relations over the sentence syntactic dependencies.

We refer now to unsupervised approaches to DE. (Reiplinger et al., 2012) benefit from hand crafted definitional patterns. Starting from a set of seed terms and patterns, term/definition pairs are iteratively acquired, together with bootstrapped new patterns. These are obtained via a generalization approach over part-of-speech and term wildcards. Additionally, two interconnected works are (De Benedictis et al., 2013) and (Faralli and Navigli, 2013), in that both bootstrap the web for acquiring large multilingual domain glossaries starting with a few seeds for term and gloss. While both systems behave similarly in extracting glosses and learning new patterns by exploiting *html* tags, they are substantially different in how acquired glosses are ranked. Specifically, the former exploits the bag-of-words representation of each extracted gloss and its intersection with the domain terminology, while the latter leverages Probabilistic Topic Models (PTM) by estimating the probability of words and term/gloss pairs to be pertinent to the domain.

3 Corpora

Our weakly supervised DE approach requires: (1) A general-domain (encyclopedic) set of seeds of textual definitions (TS) and (2) A domain-specific development set, e.g. a collection of papers (DS).

For our experiments, we use as TS the WCL Corpus (Navigli et al., 2010), a subset of Wikipedia

manually annotated with definitions and hypernyms. This dataset is constructed under the intuition that the first sentence of a Wikipedia article constitutes its textual definition. It is important to highlight that, while this dataset includes semantic information manually annotated such as definiendum or hypernym, we do not exploit any of it, which makes the seed-construction step highly flexible as it only requires the sentence definition/non-definition class. We use as DS a subset of the ACL ARC corpus (Bird et al., 2008), processed with ParsCit (Councill et al., 2008). In this dataset, a well-formedness confidence score is given to each sentence (as these come from pdf parsing and noise is introduced in the process). We exploit this information and keep 500k sentences with a score of over .95.

For evaluation, we use two datasets: The MSR-NLP [2] and the W00 corpus. The MSR-NLP is a manually constructed small list of 50 abstracts in the NLP field, amounting to 304 sentences: 49 definitions and 255 non-definitions. They are extracted from the Microsoft Academic Research website[3], where abstracts including a definition provide a "Definition Context" section. This small dataset complies with the stylistic requirements of academic abstract writing, i.e. the use of well-developed, unified, coherent and concise language, and understandability to a wide audience[4]. A different register can be found in the W00 dataset, which includes many definitional sentences that are highly domain-specific, sometimes including the definition of a very specific concept, and showing higher linguistic variability (e.g. the definiendum might not appear at the beginning of the sentence, and unlike most abstracts, citations might be present). We illustrate this difference with two sentences containing a definition from the MSR-NLP (1) and the W00 (2) corpora:

(1) The Hidden Markov Model (HMM) is a probabilistic model used widely in the fields of Bioinformatics and Speech Recognition .

(2) This corpus is collected and annotated for the GNOME project (Poesio, 2000), which aims at developing general algorithms for generating nominal expressions

Note that in the case of (2), only the sequence "GNOME project aims at developing general algorithms for generating nominal expressions" is labelled as definition in the original dataset. In this work a definitional sentence is generalized as *being* or *containing* a definition, which enables casting the task as a sentence-classification problem, which is common practice in DE (Navigli and Velardi, 2010; Boella et al., 2014; Espinosa-Anke and Saggion, 2014).

Intuitively, we would expect a general-purpose DE system to be more likely to label sentence (1), as it includes the required elements for a canonical genus-et-differentia definition. This motivates our experiments, where we attempt to fit a model iteratively to be able to perform better in sentences like (2).

4 Modelling the Data

As mentioned in Section 3, we approach the DE task as a sentence classification problem, where a sentence can be either a definition (*def*) or not (*nodef*). However, instead of modelling sentence-level features like sentence length or depth of the parse tree, we rather encode word-level features in order to exploit individual items' characteristics in terms of position within the sentence, frequency or relevance in a definition corpus. These word-level features are used for classifying each word in a sentence (*def*|*nodef*).

We adopt two extraction strategies depending on whether we operate over DS or any of the two evaluation corpora (MSR-NLP and W00). In the case DS, the goal is to extract complete high-quality definitional and non-definitional sentences. Therefore, we only consider as potential candidates for bootstrapping those sentences where all the words have the same label (i.e. discarding, for example, a 10-word sentence where nine are tagged as *def* and one as *nodef*). This is in fact the most frequent case by a large margin, so we are confident that there are very few potentially relevant sentences being left out. Since evaluation is carried out at word level, this constraint does not apply.

[2] Available at http://www.taln.upf.edu/MSR-NLP_RANLP2015

[3] http://academic.research.microsoft.com/

[4] http://www.cameron.edu/~carolynk/Abstracts.html

We exploit the potential of the Conditional Random Fields[5] algorithm (Lafferty et al., 2001) to encode prior and posterior contextual information of a given element in a sequence (in our case, a word in a sentence). Specifically, we consider a context window of [-2,2]. For each word, we generate a feature vector consisting on the following features:

1. **sur**: Surface form of the current token without stemming.
2. **lem**: Lemma of the current token.
3. **pos**: Part-of-speech of the current token.
4. **bio-np**: Whether the current word is at the beginning (B), inside (I) or outside (O) a noun phrase. Noun phrases are obtained with the following regular expression over part-of-speech tags: [JN]*N.
5. **dep**: Dependency relation between the current token and its head.
6. **head-id**: The index of the head-word (or governor) in the syntactic dependency tree.
7. **bio-def**: An extension of the *bio-np* feature that also takes into account the definition-wise position. We perform this naïvely by finding the first verb of the sentence, and tagging all words before it as definiendum and the rest as definiens. We illustrate this feature below, where each word's NP-chunking comes from the bio-np feature, *D* refers to definiendum and *d* refers to definiens.

   The⟨o-D⟩ Abwehr⟨b-D⟩ was⟨o-d⟩ a⟨o-d⟩ German⟨b-d⟩ intelligence⟨i-d⟩ organization⟨i-d⟩ from⟨o-d⟩ 1921⟨o-d⟩ to⟨o-d⟩ 1944⟨o-d⟩ .
8. **termhood**: This metric determines the importance of a candidate token to be a terminological unit by looking at its frequency in general and domain-specific corpora (Kit and Liu, 2008). It is obtained as follows:

$$\text{Termhood(w)} = \frac{\text{r}_\text{D}(\text{w})}{|\text{V}_\text{D}|} - \frac{\text{r}_\text{B}(\text{w})}{|\text{V}_\text{B}|}$$

   Where r_D is the frequency-wise ranking of word w in a domain corpus (in our case, TS), and r_B is the frequency-wise ranking of such word in a general corpus, namely the Brown corpus (Francis and Kucera, 1979). Denominators refer to the token-level size of each corpus. If word w only appears in the general corpus, we set the value of Termhood(w) to $-\infty$, and to ∞ in the opposite case.
9. **tf-gen**: Frequency of the current word in the general-domain corpus r_B (Brown Corpus).
10. **tf-dom**: Frequency of the current word in the domain-specific corpus r_D (TS).
11. **tfidf**: Tf-idf of the current word over the training set, where each sentence is considered a separate document.
12. **def_prom**: We introduce the notion of Definitional Prominence aiming at establishing the probability of a word w to appear in a definitional sentence ($s = def$). For this, we consider its frequency in definitions and non-definitions in the TS as follows:

$$\text{DefProm(w)} = \frac{\text{DF}}{|\text{Defs}|} - \frac{\text{NF}}{|\text{Nodefs}|}$$

   where $\text{DF} = \sum_{i=0}^{i=n}(s_i = \textit{def} \wedge w \in s_i)$ and $\text{NF} = \sum_{i=0}^{i=n}(s_i = \textit{nodef} \wedge w \in s_i)$. Similarly as with the *termhood* feature, in cases where a word w is only found in definitional sentences, we set the DefProm(w) value to ∞, and to $-\infty$ if it was only seen in non-definitional sentences.
13. **D_prom**: We also introduce Definiendum Prominence in order to model our intuition that a word appearing more often in position of potential *definiendum* might reveal its role as a definitional keyword. This feature is computed as follows:

$$\text{DP(w)} = \frac{\sum_{i=0}^{i=n} w_i \in \text{term}_D}{|DT|}$$

   where term_D is a noun phrase (i.e. a term candidate) appearing in potential definiendum po-

[5] We use the CRF++ toolkit: http://crfpp.googlecode.com/svn/trunk/doc/index.html

sition and |DT| refers to the size of the candidate term corpus in candidate definienda position.

14. **d_prom**: Similarly computed as D_prom, but considering position of potential definiens.

4.1 Bootstrapping

As noted in Section 3, the initial TS consists of the WCL dataset, which makes our model suitable for DE in well-formed encyclopedic texts. However, our hypothesis that it would perform poorly in a linguistically more complex setting (e.g. in a corpus like the W00 dataset) is confirmed by the results at iteration 1 (see Table 1). Our bootstrapping approach is aimed at gradually obtaining a better fit model for W00, starting from our generic baseline trained exclusively on the WCL corpus. The following description of our approach is summarized in Algorithm 1.

As mentioned above, TS is a manually labelled dataset where each sentence $s \in S$ is given a label $d \in D = \{def, nodef\}$. Likewise, DS is an unlabelled subset of the ACL-ARC corpus, which amounts to 500k sentences. The first step is to initialize (1) The training set vocabulary V, which simply contains all the words in TS; and (2) The feature set F associated to each word $w \in V$. Then, for each iteration until we reach 200, the algorithm extracts the best-scoring sentences as predicted by our CRF-based classififer (recall that only sentences where all words are assigned the same label are considered) for both labels *def* and *nodef* (s' and s'' respectively), and uses them to increase the initial feature set and vocabulary. Next, it removes s' and s'' from DS, trains and evaluates a model on both the MSR-NLP and the W00 datasets, and repeats until it reaches our manually set end point: iteration 200th.

One important aspect to consider is that increasing the size of the training data does not have an effect of the features associated to a word. Incorporating definitions having concepts related to the target domain (NLP in our case) is a step forward, but their definitional salience (expressed by def_prom, D_prom and d_prom) remains the same, as they were calculated before firing the bootstrapping algorithm. For this reason, we include a feature update step at iteration 100, our sole motivation being that, for evaluation purposes, we will have the same number of iterations before and after such step. It consists in resetting F to $\emptyset$ and recalculating it. We hypothesize that the new feature values can reflect better the linguistic idiosyncrasies of a domain-specific definitional corpus. After 200 iterations, our bootstrapped dataset TS_{boot} includes the original training data and 400 new sentences: 200 definitions and 200 non-definitions.

As the bootstrapping process advances, s' and s'' show greater linguistic variability because the training data includes more non-canonical definitions (Table 1).

Algorithm 1 Bootstrapping for DE

Require:
$TS = \{(S, d \in D)\}$ Initial labelled train seeds.
$DS = \{S\}$ Subset of the ACL-ARC corpus.
MSR-NLP: Test set 1.
W00: Test set 2.

$V := \{w : \exists\,(s,d) \in TS \wedge w \in s\}$
$F := \{f_{TS}\,(w) : w \in V\}$
1: **for** $i = 0, i < 200, i++$ **do**
  $s' = argmax_{s \in DS}\ P\,(s = def)$
  $s'' = argmax_{s \in DS}\ P\,(s = nodef)$
2: **for** $w \in s' \cup s''$ **do**
3: **if** $w \notin V$ **then**
  $F = F \cup \{f_{TS}\,(w)\}$
  $V = V \cup \{w\}$
4: **end if**
5: **end for**
  $TS = TS \cup \{(s', def)\,, (s'', nodef)\}$
  $DS = DS \setminus \{(s', def)\,, (s'', nodef)\}$
6: **if** $i = 100$ **then**
  $F = \emptyset$
7: **for** $w \in V$ **do**
  $F = F \cup \{f_{TS}\,(w)\}$
8: **end for**
9: **end if**
$model_i = trainModel\,(TS_i, F_i)$
$evaluateModel\,(model_i, \{\text{MSR-NLP, W00}\})$
10: **end for**

4.2 Post Classification Heuristics

Our last step consists in applying a post-classification heuristic inspired by (Cai et al.,

Iter	Best definition in DS	MSR-NLP			W00		
		P	R	F	P	R	F
1	A term is a word or a word sequence	100	9.09	16.68	65.38	1.25	2.47
10	An abbreviation is defined as a shortened form of a written word or phrase used in place of the full form	83.13	44.4	57.88	69.84	11.35	19.53
120	A bunsetsu is one of the linguistic units in Japanese and roughly corresponds to a basic phrase in English	25.5	90.71	39.81	60.71	69.68	64.89
182	That is to say a site is a candidate site when it is found to have either an English page linking to its Chinese version or a Chinese page linking to its English version	22.92	92.53	36.74	62.55	76.63	68.88
200	Figure 1 and Figure 2 present the overall system configuration and data flow of the integrated system	23.34	96.72	37.6	62.27	78.45	69.43

Table 1: Definitions extracted throughout the bootstrapping process from the ACL ARC corpus and P/R/F results at that iteration on the two evaluation corpora (without post-classification heuristics). Note the gradual increase in syntactic and terminological variability in the extracted definitions.

2009). It consists in a set of rules for label-switching aimed at increasing the recall and ideally without hurting precision significantly. Let w_i be a word classified as not being part of a definition (*nodef*) at iteration i, we can rectify its class (w_i^{new}) to being part of a definition (*def*) as follows:

$$w_i^{new} = \begin{cases} def & \text{if } P(w_i) = def > \theta \\ def & \text{if } P(w_i) = nodef < \lambda, w_i^{\text{syn}} = P \end{cases}$$

Where w_i^{syn} refers to the dependency relation of the word examined at iteration i, and *P* is the *predicative* syntactic function of the word.

Our goal is to increase the number of *def* words in a sentence in cases where they were discarded by a small margin. We hypothesize that this could be particularly useful in "borderline" cases (some words classified in a sentence as *def*, some as *nodef*), where this heuristics helps our algorithm to make a decision always favouring definition labelling over non-definition. As for the constants, θ and λ are empirically set to .35 and .8 respectively after experimenting with several thresholds and inspecting manually the resulting classification.

5 Evaluation

We evaluate the performance of our approach at each iteration on both datasets (MSR-NLP and W00) using the classic Precision, Recall and F-Measure scores. All the scores reported in this article are at word-level.

The learning curves shown in Figure 1 demonstrate that our approach is suitable for fitting a model to a domain-specific dataset starting from general-purpose encyclopedic seeds. Unsurprisingly, performance on the MSR-NLP corpus drops soon after reaching its peak due to the fact that the training set gradually becomes less standard. Interestingly, the feature-update step has a dramatic influence in performance in both corpora: On one hand, the performance peak in a dataset with less linguistic variability (MSR-NLP) is reached early, and after

iteration 100, where the feature update step occurs, Precision decreases, while Recall remains the same. On the other hand, the numbers in the W00 dataset are fairly stable until iteration 100, where a significant improvement in both Precision and Recall is achieved.

Let us look first at the results without applying recall-boosting post-classification heuristics: The performance of our models decreases in the MSR-NLP corpus after a few iterations (our best model is reached at iteration 23, where F=76.23), and this situation is unsurprisingly aggravated by the feature update step. However, our results improve significantly in the W00 dataset[6] after feature updating. Our best-performing model reaches F=70.72 at iteration 198.

Moreover, we observed a minor improvement after incorporating the label-switching heuristics in both corpora. Specifically, for the MSR-NLP corpus the improvement was from the aforementioned F=76.34 to F=77.46, while in the W00 dataset, it improved from F=70.72 to F=71.85. Tables 2 and 3 show Precision, Recall and F-Score for our best models in both datasets.

These numbers confirm that we are able to generate a domain and genre-sensitive model provided we have a development set available of similar characteristics. The discrepancy in terms of performance as the bootstrapping algorithm advances is an indicator that the models we obtain become more tailored towards the specific corpus, and therefore less apt for performing well in the encyclopedic genre. Our approach seems suitable for partially alleviating the lack of manually labelled domain-specific data in the DE field.

Let us also refer to the importance of having a development set as close as possible to the target corpus in terms of register and domain, and with a reasonable level of quality. In relation to this, we also performed experiments with a development set automatically constructed from the Web, but due to lack of preprocessing for noise filtering, results were unsatisfactory and therefore unreported in this paper.

As for comparative evaluation, we cannot contrast our results directly with the ones reported in (Jin et al., 2013), since while in both cases word-level evaluation is carried out, in our case we generalized all the words inside a sentence containing a definition to the label *def*. In addition, as it is pointed out in (Jin et al., 2013), only in (Reiplinger et al., 2012) there is an attempt to extract definitions from the ACL ARC corpus, but their evaluation relies on human judgement, and their reported coverage refers to a pre-defined list of terms.

[6]Note that since the W00 corpus is also a subset of the ACL ARC dataset, we first confirmed that it did not overlap with our dev-set.

	Iteration	P	R	F
Pre-PCH	198	62.69	81.11	70.72
Post-PCH	198	62.47	82.01	71.85

Table 2: Best results for the W00 dataset before (Pre-PCH) and after (Post-PCH) applying the post-classification heuristics.

	Iteration	P	R	F
Pre-PCH	23	80.69	72.24	76.23
Post-PCH	20	78.2	76.7	77.44

Table 3: Best results for both the MSR-NLP dataset before (Pre-PCH) and after (Post-PCH) applying the post-classification heuristics.

In general, the results reported in this article are consistent with the ones obtained in previous work for similar tasks. For instance, prior experiments on the WCL dataset showed results ranging from F=54.42 to F=75.16 (Navigli and Velardi, 2010; Boella et al., 2014). In the case of the W00 dataset, (Jin et al., 2013) reported numbers between F=40 and F=56 for different configurations. Since the availability of manually labelled gold standard is scarce, other authors evaluated Glossary/Definition Extraction systems in terms of manually assessed precision (Reiplinger et al., 2012; De Benedictis et al., 2013).

6 Feature Analysis

In order to understand the discriminative power of the features designed for our experiments, we computed Information Gain, which measures the decrease in entropy when the feature is present vs. ab-

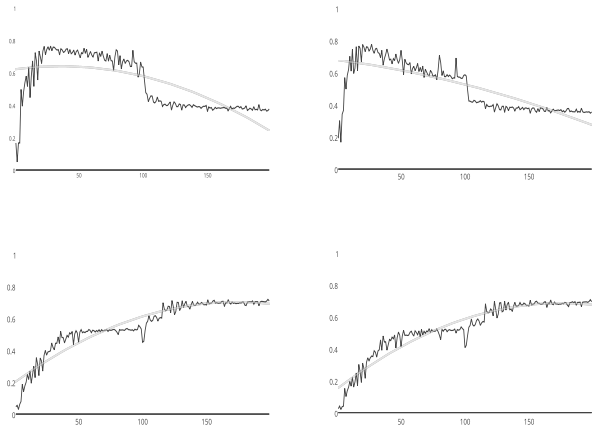

Figure 1: F-Score against iteration on the MSR-NLP (top row) and W00 datasets (bottom row), with bootstrapping + post-classification heuristics (left column) and only bootstrapping (right column).

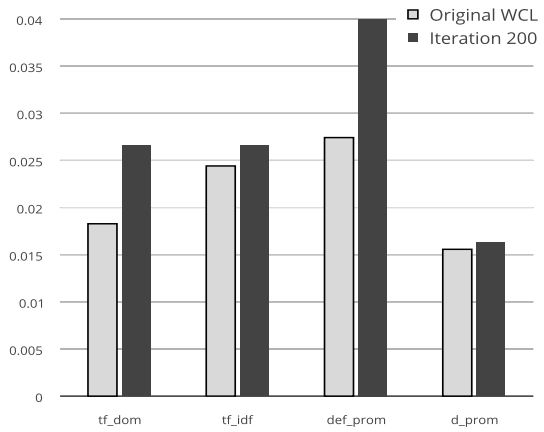

Figure 2: Information Gain for the best features at the end of the bootstrapping process. Note the substantial improvement in def_prom (definitional prominence).

sent (Forman, 2003), using the Weka toolkit (Witten and Frank, 2005). We did this for the original training set TS and the training set resulting at iteration 200 TS_{boot}. Then, we captured the top 30 features in TS_{boot}, and averaged their Information Gain score over all the available contexts. Finally, we compare these features in both datasets TS and TS_{boot} (see Figure 2).

We observe an improvement of definitionally-motivated features after iteration 100, which combined with the gradual improvement in performance in the W00 dataset, suggests that def_prom and d_prom contribute decisively to domain-specific DE, while D_prom proved less relevant. Note that in our setting, we do not focus in term/definition pairs, but rather a full-sentence definition. Therefore, we do not know a priori which term is the definiendum, and thus we do not perform a generalization step to convert it to a wildcard, which is common practice in the DE literature (Navigli and Velardi, 2010; Reiplinger et al., 2012; Jin et al., 2013; Boella et al., 2014). This provokes high sparsity in D_prom and we hypothesize that this may be the reason for this feature to not gain predictive power after many iterations or the feature update step.

7 Conclusions and Future Work

We have presented a weakly supervised DE approach that gradually increments the size of the training set with high quality definitions and clear examples of non-definitions. Two main conclusions can be drawn: (1) The definition-aware features we introduce show, in general, high informativeness for the task of DE; and (2) Our approach is valid for generating genre and domain specific training data capable of fitting corpora, even though this differs greatly in terms of content and register from the encyclopedic genre.

In addition, a small and focused benchmarking dataset of real-world definitions in the NLP domain has been released, which can be used both for linguistic and stylistic purposes and for evaluating DE systems.

These results motivate us to extend our experiments to several domains and textual genres, and to perform a longer iterative cycle where feature update is carried out more frequently. We believe that another interesting avenue for future work is multilingual definition extraction, which could benefit significantly from existing multilingual semantic networks and knowledge bases.

References

Rodrigo Alarcón, Gerardo Sierra, and Carme Bach. 2009. Description and evaluation of a definition extraction system for spanish language. In *Proceedings of the 1st Workshop on Definition Extraction*, WDE '09, pages 7–13, Stroudsburg, PA, USA. Association for Computational Linguistics.

Henning Bergenholtz and Sven Tarp. 2003. Two opposing theories: On h.e. wiegand's recent discovery of lexicographic functions. *Hermes, Journal of Linguistics*, 3-1:171–196.

Steven Bird, Robert Dale, Bonnie Dorr, Bryan Gibson, Mark Joseph, Min-Yen Kan, Dongwon Lee, Brett Powley, Dragomir Radev, and Yee Fan Tan. 2008. The acl anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC-08)*, Marrakech, Morocco, May. European Language Resources Association (ELRA). ACL Anthology Identifier: L08-1005.

Guido Boella, Luigi Di Caro, Alice Ruggeri, and Livio Robaldo. 2014. Learning from syntax generalizations for automatic semantic annotation. *Journal of Intelligent Information Systems*, pages 1–16.

Claudia Borg, Michael Rosner, and Gordon Pace. 2009. Evolutionary algorithms for definition extraction. In *Proceedings of the 1st Workshop in Definition Extraction*.

Peng Cai, HangZai Luo, and AoYing Zhou. 2009. Named entity recognition in italian using crf. In *EVALITA*.

Isaac G Councill, C Lee Giles, and Min-Yen Kan. 2008. Parscit: an open-source crf reference string parsing package. In *LREC*.

Hang Cui, Min-Yen Kan, and Tat-Seng 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*, pages 384–391. ACM.

Flavio De Benedictis, Stefano Faralli, Roberto Navigli, et al. 2013. Glossboot: Bootstrapping multilingual domain glossaries from the web. In *ACL (1)*, pages 528–538.

Lukasz Degórski, Micha Marcińczuk, and Adam Przepiórkowski. 2008. Definition extraction using a sequential combination of baseline grammars and machine learning classifiers. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco, may.

Rosa Del Gaudio, Gustavo Batista, and António Branco. 2013. Coping with highly imbalanced datasets: A case study with definition extraction in a multilingual setting. *Natural Language Engineering*, pages 1–33.

Luis Espinosa-Anke and Horacio Saggion. 2014. Applying dependency relations to definition extraction. In *Natural Language Processing and Information Systems*, pages 63–74. Springer.

Stefano Faralli and Roberto Navigli. 2013. Growing multi-domain glossaries from a few seeds using probabilistic topic models. In *EMNLP*, pages 170–181.

Tiziano Flati, Daniele Vannella, Tommaso Pasini, and Roberto Navigli. 2014. Two is bigger (and better) than one: the wikipedia bitaxonomy project. In *Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics*.

George Forman. 2003. An extensive empirical study of feature selection metrics for text classification. *The Journal of machine learning research*, 3:1289–1305.

W Nelson Francis and Henry Kucera. 1979. Brown corpus manual. *Brown University*.

Pedro Fuertes-Olivera. 2010. *Specialised Dictionaries for Learners*. Berlin/New York: De Gruyter. Lexicographica Series Maior, 136.

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, Seattle, Washington, USA, October. Association for Computational Linguistics.

Chunyu Kit and Xiaoyue Liu. 2008. Measuring mono-word termhood by rank difference via corpus comparison. *Terminology*, 14(2):204–229.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Véronique Malaisé, Pierre Zweigenbaum, and Bruno Bachimont. 2004. Detecting semantic relations between terms in definitions. In Sophia Ananadiou and Pierre Zweigenbaum, editors, *International Conference on Computational Linguistics (COLING 2004) - CompuTerm 2004: 3rd International Workshop on Computational Terminology*, pages 55–62, Geneva, Switzerland, August 29.

Ingrid Meyer. 2001. Extracting knowledge-rich contexts for terminography. *Recent advances in computational terminology*, 2:279.

Paola. Monachesi and Eline. Westerhout. 2008. What can NLP techniques do for eLearning? In *International Conference on Informatics and Systems (INFOS08)*, pages 150–156.

A Muresan and Judith Klavans. 2002. A method for automatically building and evaluating dictionary resources. In *Proceedings of the Language Resources and Evaluation Conference (LREC*.

Jun-ichi Nakamura and Makoto Nagao. 1988. Extraction of semantic information from an ordinary english dictionary and its evaluation. In *Proceedings of the 12th Conference on Computational Linguistics - Volume 2*, COLING '88, pages 459–464, Stroudsburg, PA, USA. Association for Computational Linguistics.

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 María Ruiz-Martínez. 2010. An annotated dataset for extracting definitions and hypernyms from the web. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).

Youngja Park, Roy J. Byrd, and Branimir K. Boguraev. 2002. Automatic Glossary Extraction: Beyond Terminology Identification. In *Proceedings of the 19th International Conference on Computational Linguistics*, pages 1–7, Morristown, NJ, USA. Association for Computational Linguistics.

Adam Przepiórkowski, Miroslav Spousta, Kiril Simov, Petya Osenova, Lothar Lemnitzer, Vladislav Kubo, and Beata Wójtowicz. 2007. Towards the automatic extraction of definitions in Slavic. In *Proceedings ofo the BSNLP workshop at ACL 2007*.

Josette Rebeyrolle and Ludovic Tanguy. 2000. Repérage automatique de structures linguistiques en corpus : le cas des énoncés définitoires. *Cahiers de Grammaire*, 25:153–174.

Melanie Reiplinger, Ulrich Schäfer, and Magdalena Wolska. 2012. Extracting glossary sentences from scholarly articles: A comparative evaluation of pattern bootstrapping and deep analysis. In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, pages 55–65, Jeju Island, Korea, July. Association for Computational Linguistics.

Carlos Rodríguez. 2004. *Metalinguistic Information Extraction from Specialized Texts to Enrich Computational Lexicons*. Ph.D. thesis, Universitat Pompeu Fabra.

Horacio Saggion and Robert Gaizauskas. 2004. Mining on-line sources for definition knowledge. In *17th FLAIRS*, Miami Bearch, Florida.

A. Sánchez and J. Márquez. 2005. Hacia un sistema de extracción de definiciones en textos jurídicos. In *Actas de la 1er Jornada Venezolana de Investigación en Lingüística e Informática*, pages 1–10.

Luís Sarmento, Belinda Maia, Diana Santos, Ana Pinto, and Luís Cabral. 2006. Corpógrafo V3 From Terminological Aid to Semi-automatic Knowledge Engineering. In *5th International Conference on Language Resources and Evaluation (LREC'06)*, Geneva.

Anne-Kathrin Schumann. 2011. A bilingual study of knowledge - rich context extraction in russian and german. In *Proceedings of the Fifth Language and Technology Conference*, pages 516–520.

Selja Seppälä. 2009. A proposal for a framework to evaluate feature relevance for terminographic definitions. In *Proceedings of the 1st Workshop on Definition Extraction*, WDE '09, pages 47–53, Stroudsburg, PA, USA. Association for Computational Linguistics.

Gerardo Sierra, Rodrigo Alarcón, César Aguilar, and Alberto Barrón. 2006. Towards the building of a corpus of definitional contexts. In *Proceeding of the 12th EURALEX International Congress, Torino, Italy*, pages 229–40.

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

L. Trimble. 1985. *English for Science and Technology: A Discourse Approach*. Cambridge Language Teaching Library.

Paola Velardi, Roberto Navigli, and Pierluigi D'Amadio. 2008. Mining the web to create specialized glossaries. *IEEE Intelligent Systems*, 23(5):18–25, September.

Paola Velardi, Stefano Faralli, and Roberto Navigli. 2013. Ontolearn reloaded: A graph-based algorithm for taxonomy induction. *Computational Linguistics*, 39(3):665–707.

Eline Westerhout and Paola Monachesi. 2007. Extraction of Dutch definitory contexts for elearning purposes. *Proceedings of the Computational Linguistics in the Netherlands (CLIN 2007), Nijmegen, Netherlands*, pages 219–34.

Ian H Witten and Eibe Frank. 2005. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Yulan Yan, Chikara Hashimoto, Kentaro Torisawa, Takao Kawai, Jun'ichi Kazama, and Stijn De Saeger. 2013. Minimally supervised method for multilingual paraphrase extraction from definition sentences on the web. In *HLT-NAACL*, pages 63–73.