SemEval-2015 Task 15: A Corpus Pattern Analysis Dictionary-Entry-Building Task

Vít Baisa
Masaryk University
xbaisa@fi.muni.cz

Jane Bradbury
University of Wolverhampton
J.Bradbury3@wlv.ac.uk

Silvie Cinková
Charles University
cinkova@ufal.mff.cuni.cz

Ismaïl El Maarouf
University of Wolverhampton
i.el-maarouf@wlv.ac.uk

Adam Kilgarriff
Lexical Computing Ltd
adam@sketchengine.co.uk

Octavian Popescu
IBM Research Center
o.popescu@us.ibm.com

Abstract

This paper describes the first SemEval task to explore the use of Natural Language Processing systems for building dictionary entries, in the framework of Corpus Pattern Analysis. CPA is a corpus-driven technique which provides tools and resources to identify and represent unambiguously the main semantic patterns in which words are used. Task 15 draws on the Pattern Dictionary of English Verbs (www.pdev.org.uk), for the targeted lexical entries, and on the British National Corpus for the input text.

Dictionary entry building is split into three subtasks which all start from the same concordance sample: 1) CPA parsing, where arguments and their syntactic and semantic categories have to be identified, 2) CPA clustering, in which sentences with similar patterns have to be clustered and 3) CPA automatic lexicography where the structure of patterns have to be constructed automatically.

Subtask 1 attracted 3 teams, though none could beat the baseline (rule-based system). Subtask 2 attracted 2 teams, one of which beat the baseline (majority-class classifier). Subtask 3 did not attract any participant.

The task has produced a major semantic multidataset resource which includes data for 121 verbs and about 17,000 annotated sentences, and which is freely accessible.

1 Introduction

It is a central vision of NLP to represent the meanings of texts in a formalised way, amenable to automated reasoning. Since its birth, SEMEVAL (or SENSEVAL as it was then; (Kilgarriff and Palmer, 2000)) has been part of the programme of enriching NLP analyses of text so they get ever closer to a 'meaning representation'. In relation to lexical information, this meant finding a lexical resource which

- identified the different meanings of words in a way that made high-quality disambiguation possible,
- represented those meanings in ways that were useful for the next steps of building meaning representations.

Most lexical resources explored to date have had only limited success, on either front. The most obvious candidates—published dictionaries and WordNets—look like they might support the first task, but are very limited in what they offer to the second.

FrameNet moved the game forward a stage. Here was a framework with a convincing account of how the lexical entry might contribute to building the meaning of the sentence, and with enough meat in the lexical entries (e.g. the verb frames) so that it might support disambiguation. Papers such as (Gildea and Jurafsky, 2002) looked promising, and in 2007 there was a SEMEVAL task on Frame Semantic Structure Extraction (Baker et al., 2007) and in 2010, one on Linking Events and Their Participants (Ruppenhofer et al., 2010).

While there has been a substantial amount of follow-up work, there are some aspects of FrameNet that make it a hard target.

- It is organised around frames, rather than words, so inevitably its priority is to give a co-

Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pages 315–324,
Denver, Colorado, June 4-5, 2015. ©2015 Association for Computational Linguistics

herent account of the different verb senses in a frame, rather than the different senses of an individual verb. This will tend to make it less good for supporting disambiguation.

- Frames are not 'data-driven': they are the work of a theorist (Fillmore) doing his best to make sense of the data for a set of verbs. The prospects of data-driven frame discovery are, correspondingly, slim.
- While FrameNet has worked hard at being systematic in its use of corpus data, FrameNetters looked only for examples showing the verb being used in the relevant sense. From the point of view of a process that could possibly be automated, this is problematic.

An approach which bears many similarities to FrameNet, but which starts from the verb rather than the frame, and is more thoroughgoing in its empiricism, is Hanks's Corpus Pattern Analysis (Hanks and Pustejovsky, 2005; Hanks, 2012; Hanks, 2013).

2 Corpus Pattern Analysis

Corpus Pattern Analysis (CPA) is a new technique of language analysis, which produces the main patterns of use of words in text. Figure 1 is a sample lexical entry from the main output of CPA, the Pattern Dictionary of English Verbs[1] (PDEV).

This tells us that, for the verb *abolish*, three patterns were found. For each pattern it tells us the percentage of the data that it accounted for, its grammatical structure and the semantic type (drawn from a shallow ontology of 225 semantic types[2]) of each of the arguments in this structure. For instance, pattern 1 means: i) that the subject is preferably a word referring to [[Human]] or [[Institution]] (semantic alternation), and ii) that the object is preferably [[Action]], [[Rule]] or [[Privilege]].

It also tells us the implicature (which is similar to a "definition" in a traditional dictionary) of a sentence exemplifying the pattern: that is, if we have a sentence of the pattern **[[Institution | Human]]** `abolish` **[[Action=Punishment | Rule | Privilege]]**, then we know that **[[Institution | Human]]** `formally declares that` **[[Action=Punishment | Rule | Privilege]]** `is no longer legal or operative`. *Abolish* has only one sense. For many verbs, there will be multiple senses, each with one or more pattern.

There are currently full CPA entries for more than 1,000 verbs with a total of over 4,000 patterns. For each verb a random sample of (by default) 250 corpus instances was examined, used to build the lexical entry, and tagged with the senses and patterns they represented. For commoner verbs, more corpus lines were examined. The corpus instances were drawn from the written part of the British National Corpus[3] (BNC).

PDEV has been studied from different NLP perspectives, all mainly involved with Word Sense Disambiguation and semantic analysis (Cinková et al., 2012a; Holub et al., 2012; El Maarouf et al., ; El Maarouf and Baisa, 2013; Kawahara et al., ; Popescu, 2013; Popescu et al., ; Pustejovsky et al., 2004; Rumshisky et al.,). For example, (Popescu, 2013) described experiments in modeling finite state automata on a set of 721 verbs taken from PDEV. The author reports an accuracy of over 70% in pattern disambiguation. (Holub et al., 2012) trained several statistical classifiers on a modified subset of 30 PDEV entries (Cinková et al., 2012c) using morpho-syntactic as well as semantic features, and obtained over 80% accuracy. On a smaller set of 20 high frequency verbs (El Maarouf and Baisa, 2013) reached a similar 0.81 overall F1 score with a supervised SVM classifier based on dependency parsing and named entity recognition features.

The goal of Task 15 at SemEval 2015 are i) to explore in more depth the mechanics of corpus-based semantic analysis and ii) to provide a high-quality standard dataset as well as baselines for the advancement of semantic processing. Given the complexity and wealth of PDEV, a major issue was to select relevant subtasks and subsets. The task was eventually split into three essential steps in building a CPA lexical entry, that systems could tackle separately:

1. *CPA parsing*: all sentences in the dataset to be syntactically and semantically parsed.
2. *CPA clustering*: all sentences in the dataset to be grouped according to their similarities.

[1] `http://pdev.org.uk`

[2] `http://pdev.org.uk/#onto`

[3] `http://www.natcorp.ox.ac.uk/`

1	*Pattern* *Implicature*	**Institution or Human abolishes Action or Rule or Privilege** Institution or Human formally declares that Action = Punishment or Rule or Privilege is no longer legal or operative	58.8%
2	*Pattern* *Implicature*	**Institution 1 or Human abolishes Institution 2 or Human_Role** Institution 1 or Human formally puts an end to Institution 2 or Human_Role	24.4%
3	*Pattern* *Implicature*	**Process abolishes State_of_Affairs** Process brings State_of_Affairs to an end	14.4%

Figure 1: PDEV Entry for *abolish*.

Tag	Definition
subj	Subject
obj	Object
iobj	Indirect Object
advprep	Adverbial Preposition or other Adverbial/Verbal Link
acomp	Adverbial or Verb Complement
scomp	Noun or Adjective complement

Table 1: Syntactic tagset used for subtask 1.

3. *CPA lexicography*: all verb patterns found in the dataset to be described in terms of their syntactic and semantic properties.

3 Task Description

In order to encourage participants to design systems which could successfully tackle all three subtasks, all tasks were to be evaluated on the same set of verbs. As opposed to previous experiments on PDEV, it was decided that the set of verbs from the test dataset would be different from the set of verbs given in the training set. This was meant to avoid limiting tasks to supervised approaches and to encourage innovative approaches, maybe using patterns learnt in an unsupervised manner from very large corpora and other resources. This also implied that the dataset would be constructed so as to make it possible for systems to generalize from the behaviour and description of one set of verbs to a set of unseen verbs used in similar structures, as human language learners do. Although this obviously makes the task harder, it was hoped that this would put us in a better position to evaluate current limits of automatic semantic analysis.

3.1 Subtask 1: CPA Parsing

The CPA parsing subtask focuses on the detection and classification (syntactic and semantic) of the arguments of the verb. The subtask is similar to Semantic Role Labelling (Carreras and Marquez, 2004) that arguments will be identified in the dependency parsing paradigm (Buchholz and Marsi, 2006), using head words instead of phrases.

The syntactic tagset was designed specially for this subtask and kept to a minimum, and the semantic tagset was based on the CPA Semantic Ontology.

In Example (1), this would mean identifying *government* as subject of *abolish*, from the [[Institution]] type, and *tax* as object belonging to [[Rule]]. The expected output is represented in XML format in Example (2).

(1) *In 1981 the Conservative government abolished capital transfer tax capital transfer tax and replaced it with inheritance tax.*

(2) *In 1981 the Conservative* <entity syn=‘subj’ sem=‘Institution’> **government** </entity> <entity syn=‘v’ sem=‘-’> **abolished** </entity> *capital transfer* <entity synt=‘obj’ sem=‘Rule’> **tax** </entity> *capital transfer tax and replaced it with inheritance tax*

The only dependency relations shown are those involving the node verb. Thus, for example, the dependency relation between *Conservative* and *government* is not shown. Also only the relations in Table 1 are shown. The relation between *abolished* and *replaced* is not shown as it is not one of the targeted dependency relations. The input text consisted of individual sentences one word per line with both ID and FORM fields, and in which only the target verb token was pre-tagged.

3.2 Subtask 2: CPA Clustering

The CPA clustering subtask is similar to a Word Sense Discrimination task in which systems have to

Layer	Annotator	dataset	observations	categories	Kappa (Cohen)	F-score
Syn	Annotator 1	both	3,662	5	0.898	0.924
	Annotator 2	train	4,106	5	0.752	0.789
	Annotator 3	test	1,518	5	0.931	0.942
Sem	Annotator 1	both	3,662	108	0.649	0.693
	Annotator 2	train	4,106	113	0.444	0.498
	Annotator 3	test	1,518	75	0.765	0.782

Table 2: Inter-annotator figures where annotators are compared to the expert (annotator 4) who reviewed all the annotations (Microcheck Task 1).

predict which pattern a verb instance belongs to.

With respect to *abolish* (Figure 1), it would involve identifying all sentences containing the verb *abolish* which belonged to the same pattern (one of the patterns in Figure 1) and tagging them with the same number.

3.3 Subtask 3: CPA Automatic Lexicography

The CPA automatic lexicography subtask aims to evaluate how systems can approach the design of a lexicographical entry within CPA's framework.

The input was, as for the other tasks, plain text with node verb identified. The output format was a variant of that shown in Figure 1, simplified to a form which would be more tractable by systems while still being a relevant representation from the lexicographical perspective.

Specifically, contextual roles were discarded and semantic alternations were decomposed into semantic strings[4] so that pattern 1 in Figure 1 would give rise to six strings (with V for the verb, here *abolish*):

```
[[Human]] V [[Action]]
[[Human]] V [[Rule]]
[[Human]] V [[Privilege]]
[[Institution]] V [[Action]]
[[Institution]] V [[Rule]]
[[Institution]] V [[Privilege]]
```

This transformation from the PDEV format as in Figure 1 was done automatically and checked manually. These strings are different to (and generally more numerous than) the patterns evaluated in subtask 2. The goal of this subtask was to generalize sentence examples for each verb and create a list of possible semantic strings. This subtask was autonomous with respect to other subtasks in that participants did not have to return the set of sentences which matched their candidate patterns, patterns were evaluated independently.

4 Task Data

4.1 The Microcheck and Wingspread Datasets

All subtasks (except the first) include two setups and their associated datasets: the number of patterns for each verb is disclosed in the first dataset but not in the second. This setup was created to see whether it would influence the results.

The two datasets were also created in the hope that system development would start on the first small and carefully crafted dataset (Microcheck) and only then be tested on a larger and more varied subset of verbs (Wingspread)[5].

4.2 Annotation Process

Both Microcheck and Wingspread start from data extracted from PDEV and the manually pattern-tagged BNC. We took only verbs declared as complete and started by the same lexicographer, so that each verb had been checked twice: once by the lexicographer who compiled the entry and once by the editor-in-chief. Some tagging errors may have slipped in but the tagging is generally of high quality (Cinková et al., 2012a; Cinková et al., 2012b). Additional checks have been performed on Microcheck, since this was the dataset chosen for subtask 1, for which data had to be created. This section describes the annotation process.

PDEV contains only one kind of link between a given pattern and a given corpus instance: each verb token found in the sample is tagged with a pattern identifier, and the pattern then specifies syntactic

[4] See (Bradbury and El Maarouf, 2013).

[5] The datasets as well as the systems' outputs will soon be made publicly available on the task website.

V	P	I	IMP	%MP	V	P	I	IMP	%MP
boo	2	36	27	0.769	ascertain	2	7	4	0.676
teeter	2	28	23	0.828	totter	2	19	12	0.697
begrudge	2	19	11	0.678	tense	3	37	23	0.628
avert	2	240	230	0.958	belch	3	24	14	0.612
breeze	2	12	7	0.679	attain	3	240	200	0.833
wing	2	22	19	0.867	avoid	3	242	176	0.728
brag	2	29	18	0.692	adapt	4	182	98	0.583
sue	2	247	242	0.980	advise	8	230	84	0.391
bluff	2	25	14	0.673	ask	9	573	299	0.518
afflict	2	179	172	0.961	SUM	59	2,423	1,689	—
bludgeon	2	32	16	0.667	AVERAGE	2.95	121.15	84.45	0.721

Table 3: Statistics on the Wingspread test dataset with V standing for verb, P for patterns, I for instances, IMP for instances of majority pattern, and %MP for proportion of the majority pattern.

V	P	I	IMP	%MP	V	P	I	IMP	%MP
appreciate	2	160	215	0.765	apprehend	3	77	123	0.652
crush	5	62	170	0.413	decline	3	135	201	0.690
continue	7	71	203	0.401					
undertake	2	204	228	0.896	SUM	30	749	1,280	—
operate	8	40	140	0.300	AVERAGE	4.286	107	182.857	0.588

Table 4: Statistics on the Microcheck test dataset; abbreviations as for previous table.

roles and their semantic types. The job in subtask 1 annotation consists of tagging the arguments of each token in the sample, both syntactically and semantically (see Table 1 for tagsets of each layer). The syntactic information was the same as for subtask 3 except that category names were shortened and pairs of categories were merged in two places.[6]

The annotation was carried out by 4 annotators, with 3 for the training data and 3 for test data, and 2 annotators annotating both training and test data, one of them being an expert PDEV annotator. Annotators could ask for feedback on the task at any moment, and any doubts were cleared by the expert annotator. Each pair of annotators annotated one share of the dataset, and their annotation was double-checked by the expert annotator. The agreement was not very high (e.g. Annotator 2, see Table 2) in some cases so the double-check by the expert annotator was crucial. Table 2 reports the agreement in terms of F-score and Cohen's Kappa (Cohen, 1960) between each annotator and the expert annotator.[7]

4.3 Statistics on the Data

Strict rules were implemented to develop a high-quality and consistent dataset:

1. PDEV patterns discriminate exploited[8] uses of a pattern using a different tag; these were left aside for the CPA task.
2. For the test set, when patterns contained at least one semantic type or grammatical category which was not covered in the training set, they were discarded.
3. Only patterns which contained more than 3 examples were kept in the final dataset.

Applying these filters led to the Microcheck dataset, containing 28 verbs (train: 21; test: 7), 378 patterns (train: 306; test: 72) with 4,529 annotated sentences (train: 3,249; test: 1,280) and to the Wingspread dataset set containing 93 verbs (train:

[6] See `http://alt.qcri.org/semeval2015/task15/index.php?id=appendices`

[7] The expert did not start from scratch, but from other annotators' work. Since his target was the conformity of the tagging with guidelines as well as with CPA's principles, we maintain that the expert would have produced a very similar output had he not started from the product of other annotators, who themselves used the output of a system to speed up their work.

[8] An exploitation corresponds to an anomalous use of a pattern, as in a figurative use.

73; test: 20), 856 patterns (train: 652; test: 204), and 12,440 annotated sentences (train: 10,017; test: 2,423). More detailed figures for the test datasets are provided in Tables 3 and 4.

4.4 Metrics

The final score for all subtasks is the average of F-scores over all verbs (Eq. 1). What varies across subtasks is the way Precision and Recall are defined.

$$\begin{aligned} \mathrm{F1}_{\mathrm{verb}} &= \frac{2 \times \mathrm{Precision}_{\mathrm{verb}} \times \mathrm{Recall}_{\mathrm{verb}}}{\mathrm{Precision}_{\mathrm{verb}} + \mathrm{Recall}_{\mathrm{verb}}} \\ \mathrm{Score}_{\mathrm{Task}} &= \frac{\sum_{i=1}^{n_{\mathrm{verb}}} \mathrm{F1}_{\mathrm{verb}_i}}{n_{\mathrm{verb}}} \end{aligned} \tag{1}$$

Subtask 1. Equation 2 illustrates that Precision and Recall are computed on all tags, both syntactic and semantic. To count as correct, tags had to be set on the same token as in the gold standard.

$$\begin{aligned} \mathrm{Precision} &= \frac{\text{Correct tags}}{\text{Retrieved tags}} \\ \mathrm{Recall} &= \frac{\text{Correct tags}}{\text{Reference tags}} \end{aligned} \tag{2}$$

Subtask 2. Clustering is known to be difficult to evaluate. Subtask 2 used the B-cubed definition of Precision and Recall, first used for coreference (Bagga and Baldwin, 1999) and later extended to cluster evaluation (Amigó et al., 2009). Both measures are averages of the precision and recall over all instances. To calculate the precision of each instance we count all correct pairs associated with this instance and divide by the number of actual pairs in the candidate cluster that the instance belongs to. Recall is computed by interchanging Gold and Candidate clusterings (Eq. 3).

$$\begin{aligned} \mathrm{Precision}_{\mathrm{i}} &= \frac{\mathrm{Pairs}_{\mathrm{i}} \text{ in Candidate found in Gold}}{\mathrm{Pairs}_{\mathrm{i}} \text{ in Candidate}} \\ \mathrm{Recall}_{\mathrm{i}} &= \frac{\mathrm{Pairs}_{\mathrm{i}} \text{ in Gold found in Candidate}}{\mathrm{Pairs}_{\mathrm{i}} \text{ in Gold}} \end{aligned} \tag{3}$$

Subtask 3. This task was evaluated as a slot-filling exercise (Makhoul et al., 1999), so the scores were computed by taking into account the kinds of errors that systems make over the 9 slots: errors of Insertion, Substitution, Deletion. Equation 4 formulates how Precision and Recall are computed.

$$\begin{aligned} \mathrm{Precision} &= \frac{\mathrm{Correct}}{\mathrm{Correct} + \mathrm{Subst} + \mathrm{Ins}} \\ \mathrm{Recall} &= \frac{\mathrm{Correct}}{\mathrm{Correct} + \mathrm{Subst} + \mathrm{Del}} \end{aligned} \tag{4}$$

In order not to penalize systems, the best match was computed for each Candidate pattern, and one candidate pattern could match more than one Gold pattern. When a given slot was filled both in the Gold data and the Candidate data, this counted as a "match". When not, it was a Deletion. If a slot was filled in the run but not in the gold, it was counted as an Insertion. When a match (aligned slots) was also a semantic type match, it was Correct (1 point). When not, it was a Substitution; the CPA ontology was used to allow for partial matches, allowing hypernyms and hyponyms. For that particular task, the maximum number of Candidate patterns was limited to 150% with respect to the number in the Gold set.

5 Evaluation

The evaluation was split into 2 phases (one week for each): a feedback phase and a validation phase. The reason for this was to allow for the detection of unforeseen issues in the output of participants' systems so as to prepare for any major problem. However, this was not put to use by participants since only one team submitted their output in the first phase which also happened to be their final submission.

5 teams[9] participated in the task, but none participated in more than one subtask. Subtask 1 attracted 3 teams and subtask 2 attracted 2, while subtask 3 did not receive any submissions. Systems were allowed 3 runs on each subtask and each dataset, and were asked to indicate which would be the official one. The following subsections report in brief on the main features of their systems (for more details see relevant papers in SemEval proceedings).

5.1 Subtask 1

All systems for this subtask used syntactic dependencies and named entities as features. Since the

[9] Unfortunately, teams BOB90 and FANTASY did not submit articles, so it is difficult to analyze their results.

Category	#Gold	CMILLS	FANTASY	BLCUNLP	baseline
subj	1,008	0.564	0.694	0.739	**0.815**
obj	777	0.659	**0.792**	0.777	0.783
Human	580	0.593	**0.770**	0.691	0.724
Activity	438	0.450	**0.479**	0.393	0.408
acomp	308	0.545	0.418	0.702	**0.729**
LexicalItem	303	0.668	**0.830**	0.771	0.811
advprep	289	0.621	0.517	0.736	**0.845**
State Of Affairs	192	**0.410**	0.276	0.373	0.211
Institution	182	0.441	**0.531**	0.483	0.461
Action	115	0.421	**0.594**	0.526	0.506

Table 6: Detailed scores for subtask 1 (10 most frequent categories).

Team	Score
baseline	0.624
FANTASY	**0.589**
BLCUNLP	0.530
CMILLS	0.516

Table 5: Official scores for subtask 1.

subtask allowed it, some systems used external resources such as Wordnet or larger corpora.

BLCUNLP (Feng et al., 2015) used the Stanford CoreNLP package[10] to get POS, NE and basic dependency features. These features were used to predict both syntax and semantic information. The method did not involve the use of a statistical classifier.

CMILLS (Mills and Levow, 2015) used three models to solve the task: one for argument detection, and the other two for each layer. Argument detection and syntactic tagging were performed using a MaxEnt supervised classifier, while the last was based on heuristics. CMILLS also reported the use of an external resource, the enTentTen12 (Jakubíček et al., 2013) corpus available in Sketch Engine (Kilgarriff et al., 2014).

FANTASY approached the subtask in a supervised setting to predict first the syntactic tags, and then the semantic tags. The team used features from the MST parser[11], as well as Stanford CoreNLP for NE, Wordnet[12], they also applied word embedding representations to predict the output of each layer.

The baseline system was a rule-based system taking as input the output of the BLLIP parser (Charniak and Johnson, 2005), and mapping heads of relevant dependency relations to the most probable tags from subtask 1 tagset. The semantic tags were only then added to those headwords based on the most frequent semantic category found in the training set.

5.2 Subtask 2

As opposed to subtask 1, systems in subtask 2 used very few semantic and syntactic resources.

BOB90 used a supervised approach to tackle the clustering problem. The main features used were preposition analyses.

DULUTH (Pedersen, 2015) used an unsupervised approach and focused on lexical similarity (both first and second order representations) based on unigrams and bigrams (see SenseClusters[13]). The number of clusters was predicted on the basis of the best value for the clustering criterion function. The team also performed some corpus pre-processing, like conversion to lower case and conversion of all numeric values to a string.

The baseline system clusters everything together, so its score depends on the distribution of patterns: the more a pattern covers all instances of the data (majority class), the higher the baseline score.

6 Results

6.1 Subtask 1

As previously noted, subtask 1 provided only one dataset, Microcheck. The results on the test set are

[10] http://nlp.stanford.edu/software/corenlp.shtml

[11] http://www.seas.upenn.edu/~strctlrn/MSTParser/MSTParser.html

[12] http://wordnet.princeton.edu/

[13] http://senseclusters.sourceforge.net

described in Table 5: FANTASY is the best system with 0.589 average F1 score, but does not beat the baseline (0.624).

It is worth noting that, on the same set of verbs, BLCUNLP and FANTASY are almost on a par, but since the former did not submit one verb file, the score gap is more significant. FANTASY is a more precise system while BLCUNLP has higher recall.

To get a better picture of the results, Table 6 provides the F-scores for the ten most frequent categories in the test set. We can see that FANTASY has the best semantic model since it gets the highest scores on most semantic categories (except for *State Of Affairs*) and systematically beats the baseline, which assigns a word the most frequent semantic category in the training set. The baseline and BCUNLP however get higher scores on most syntactic relations except on *obj*, where the difference is low. The gap is much more significant on *advprep* and *acomp*, which suggests that FANTASY does not properly handle prepositional complements correctly (and/or causal complements). This could be due to the choice of parser or to model parameters. Overall, it seems that progress can still be made, since systems can benefit from one another.

6.2 Subtask 2

Subtask 2 was evaluated on both datasets. BOB90 only submitted one run while DULUTH submitted three. The results are displayed on Table 7. For this task, only BOB90 beat the baseline with a higher amplitude on Microcheck (+0.153) than on Wingspread (+0.071). This high score welcomes a more detailed evaluation of the system, since it would seem that, as also found for subtask 1, prepositions play a substantial role in CPA patterns and semantic similarity.

It can also be observed that overall results are better on Wingspread. This seems to be mainly due to the higher number of verbs with a large majority class in Wingspread (see Table 3), since the baseline system scores 0.72 on Wingspread, and 0.588 on Microcheck. This shows that when the distribution of patterns is highly skewed, the evaluation of systems is difficult, and tends to underrate potentially useful systems.

Team	Scores	
	Microcheck	Wingspread
BOB90	**0.741**	**0.791**
baseline	0.588	0.720
DULUTH-1 (off)	0.525	0.604
DULUTH-2	0.439	0.581
DULUTH-3	0.439	0.615

Table 7: Official scores for subtask 2.

7 Conclusion

This paper introduces a new SemEval task to explore the use of Natural Language Processing systems for building dictionary entries, in the framework of Corpus Pattern Analysis. Dictionary entry building is split into three subtasks: 1) CPA parsing, where arguments and their syntactic and semantic categories have to be identified, 2) CPA clustering, in which sentences with similar patterns have to be clustered and 3) CPA automatic lexicography where the structure of patterns have to be constructed automatically.

Drawing from the Pattern Dictionary of English Verbs, we have produced a high-quality resource for the advancement of semantic processing: it contains 121 verbs connected to a corpus of 17,000 sentences. This resource will be made freely accessible from the task website for more in depth future research.

Task 15 has attracted 5 participants, 3 on subtask 1 and 2 on subtask 2. Subtask 1 proved to be more difficult for participants than expected, since no system beat the baseline. We however show that the submissions possess interesting features that should be put to use in future experiments on the dataset. Subtask 2's baseline was beaten by one of the participants on a large margin, despite the fact that the baseline is very competitive.

It seems that splitting the task into 3 subtasks has had the benefit of attracting different approaches (supervised and unsupervised) towards the common target of the task, which is to build a dictionary entry. Lexicography is such a complex task that it needs major efforts from the NLP community to support it. We hope that this task will stimulate more research and the development of new approaches to the automatic creation of lexical resources.

Acknowledgments

We are very grateful for feedback on the task from Ken Litkowski as well as from participants who greatly contributed to the overall quality of the task. We would also like to thank SemEval's organizers for their support. The work was also supported by the UK's AHRC grant [DVC, AH/J005940/1, 2012-2015], by the Ministry of Education of Czech Republic within the LINDAT-Clarin project LM2010013, by the Czech-Norwegian Research Programme within the HaBiT Project 7F14047 and by the Czech Science Foundation grant [GA15-20031S].

References

Enrique Amigó, Julio Gonzalo, Javier Artiles, and Felisa Verdejo. 2009. A Comparison of Extrinsic Clustering Evaluation Metrics Based on Formal Constraints. *Information Retrieval*, 12(4):461–486.

Amit Bagga and Breck Baldwin. 1999. Cross-document event coreference: Annotations, experiments, and observations. In *Proceedings of the Workshop on Coreference and its Applications*, pages 1–8.

Collin Baker, Michael Ellsworth, and Katrin Erk. 2007. SemEval-2007 Task 19: Frame Semantic Structure Extraction. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 99–104, Prague, Czech Republic.

Jane Bradbury and Ismaïl El Maarouf. 2013. An empirical classification of verbs based on Semantic Types: the case of the 'poison' verbs. In *Proceedings of JSSP2013*, pages 70–74, Trento,Italy.

Sabine Buchholz and Erwin Marsi. 2006. CoNLL-X shared task on multilingual dependency parsing. In *Proceedings of CoNLL*, New York, USA.

Xavier Carreras and Lluis Marquez. 2004. Introduction to the CoNLL-2004 shared task: Semantic role labeling. In *Proceedings of CoNLL*, Boston, USA.

Eugene Charniak and Mark Johnson. 2005. Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 173–180.

Silvie Cinková, Martin Holub, and Vincent Kríž. 2012a. Managing Uncertainty in Semantic Tagging. In *Proceedings of 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 840–850, Avignon, France.

Silvie Cinková, Martin Holub, and Vincent Kríž. 2012b. Optimizing semantic granularity for NLP - report on a lexicographic experiment. In *Proceedings of the 15th EURALEX International Congress*, pages 523–531, Oslo, Norway.

Silvie Cinková, Martin Holub, Adam Rambousek, and Lenka Smejkalová. 2012c. A database of semantic clusters of verb usages. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, pages 3176–3183, Istanbul, Turkey.

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46.

Ismaïl El Maarouf and Vít Baisa. 2013. Automatic classification of semantic patterns from the Pattern Dictionary of English Verbs. In *Proceedings of JSSP2013*, pages 95–99, Trento, Italy.

Ismaïl El Maarouf, Jane Bradbury, Vít Baisa, and Patrick Hanks. Disambiguating Verbs by Collocation: Corpus Lexicography meets Natural Language Processing. In *Proceedings of LREC*, pages 1001–1006, Reykjavik, Iceland.

Yukun Feng, Qiao Deng, and Dong Yu. 2015. BLCUNLP: Corpus Pattern Analysis for Verbs Based on Dependency Chain. In *Proceedings of SemEval 2015*, Denver, USA.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic Labeling of Semantic Roles. *Computational Linguistics*, 28(3):245–288, September.

Patrick Hanks and James Pustejovsky. 2005. A Pattern Dictionary for Natural Language Processing. *Revue Française de linguistique applique*, 10:2.

Patrick Hanks. 2012. How people use words to make meanings: Semantic types meet valencies. In A. Boulton and J. Thomas, editors, *Input, Process and Product: Developments in Teaching and Language Corpora*, pages 54–69. Brno.

Patrick Hanks. 2013. *Lexical Analysis: Norms and Exploitations*. MIT Press, Cambridge, MA.

Martin Holub, Vincent Kríž, Silvie Cinková, and Eckhard Bick. 2012. Tailored Feature Extraction for Lexical Disambiguation of English Verbs Based on Corpus Pattern Analysis. In *Proceedings of the 24th International Conference on Computational Linguistics (Coling 2012)*, pages 1195–1209, Mumbai, India.

Miloš Jakubíček, Adam Kilgarriff, Vojtěch Kovář, Pavel Rychlý, and Vít Suchomel. 2013. The TenTen Corpus Family. In *Proceedings of the International Conference on Corpus Linguistics*.

Daisuke Kawahara, Daniel W Peterson, Octavian Popescu, and Martha Palmer. Inducing Example-based Semantic Frames from a Massive Amount of Verb Uses. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 58–67.

Adam Kilgarriff and Martha Palmer. 2000. Introduction to the Special Issue on SENSEVAL. *Computers and the Humanities*, 34:1–2.

Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. The Sketch Engine: ten years on. *Lexicography*, 1(1):7–36.

John Makhoul, Francis Kubala, Richard Schwartz, and Ralph Weischedel. 1999. Performance Measures For Information Extraction. In *Proceedings of DARPA Broadcast News Workshop*, pages 249–252.

Chad Mills and Gina-Anne Levow. 2015. CMILLS: Adapting SRL Features to Dependency Parsing. In *Proceedings of SemEval 2015*, Denver, USA.

Ted Pedersen. 2015. Duluth: Word Sense Discrimination in the Service of Lexicography. In *Proceedings of SemEval 2015*, Denver, USA.

Octavian Popescu, Martha Palmer, and Patrick Hanks. Mapping CPA onto OntoNotes Senses. In *Proceedings of LREC*, pages 882–889, Reykjavik, Iceland.

Octavian Popescu. 2013. Learning corpus patterns using finite state automata. In *Proceedings of the 10th International Conference on Computational Semantics*, pages 191–203, Potsdam, Germany.

James Pustejovsky, Patrick Hanks, and Anna Rumshisky. 2004. Automated Induction of Sense in Context. In *Proceedings of COLING*, Geneva, Switzerland.

Anna Rumshisky, Patrick Hanks, Catherine Havasi, and James Pustejovsky. Constructing a corpus-based ontology using model bias. In *Proceedings of FLAIRS*, pages 327–332, Melbourne, FL.

Josef Ruppenhofer, Caroline Sporleder, Roser Morante, Collin Baker, and Martha Palmer. 2010. SemEval-2010 Task 10: Linking Events and Their Participants in Discourse. In *5th International Workshop on Semantic Evaluation*, pages 45–50, Uppsala, Sweden.