SemEval-2016 Task 14: Semantic Taxonomy Enrichment

David Jurgens
Stanford University
jurgens@cs.stanford.edu

Mohammad Taher Pilehvar
University of Cambridge
mp792@cam.ac.uk

Abstract

Manually constructed taxonomies provide a crucial resource for many NLP technologies, yet these resources are often limited in their lexical coverage due to their construction procedure. While multiple approaches have been proposed to enrich such taxonomies with new concepts, these techniques are typically evaluated by measuring the accuracy at identifying relationships between words, e.g., that a dog is a canine, rather relationships between specific concepts. Task 14 provides an evaluation framework for automatic taxonomy enrichment techniques by measuring the placement of a new concept into an existing taxonomy: Given a new word and its definition, systems were asked to attach or merge the concept into an existing WordNet concept. Five teams submitted 13 systems to the task, all of which were able to improve over the random baseline system. However, only one participating system outperformed the second, more-competitive baseline that attaches a new term to the first word in its gloss with the appropriate part of speech, which indicates that techniques must be adapted to exploit the structure of glosses.

1 Introduction

Semantic networks and ontologies are key resources in Natural Language Processing. Of these resources, WordNet (Fellbaum, 1998), the de facto standard lexical database of English, has remained in widespread use over the past two decades, with a broad range of applications such as Word Sense Disambiguation (Navigli, 2009), Query expansion and Information Retrieval (Varelas et al., 2005; Fang, 2008), sentiment analysis (Esuli and Sebastiani, 2006), and semantic similarity measurement (Budanitsky and Hirst, 2006a; Pilehvar et al., 2013). The performances of these WordNet-based techniques are directly affected by the lexical coverage of WordNet's vocabulary, especially if applied to specific domains and social media texts. However, the manual maintenance of WordNet is an expensive endeavour which requires significant effort and time. As a result, WordNet is not updated frequently and omits many lemmas and senses, such as those from domain specific lexicons (e.g., DNA replication, regular expression, and long shot), creative slang usages (e.g., homewrecker), or those for technology or entities that came into recent existence (e.g., selfie, mp3).

Hence, a variety of techniques have tried to tackle the coverage limitation of WordNet, often by drawing new word senses from other domain-specific or collaboratively-constructed dictionaries and adding the new word senses to the WordNet hierarchy (Poprat et al., 2008; Snow et al., 2006; Toral et al., 2008; Yamada et al., 2011; Jurgens and Pilehvar, 2015). However, these approaches have usually been tested on relatively small datasets, often testing for word-level relationships without precisely measuring integration accuracy at the concept level. Similarly, other techniques have been proposed for automatically discovering novel senses of words (Lau et al., 2012); however, these senses were not re-integrated into the taxonomy.

Given the availability of large-scale dictionaries such as Wiktionary, Task 14 is designed to inspire

Proceedings of SemEval-2016, pages 1092–1102,
San Diego, California, June 16-17, 2016.

new automated approaches for using the definitions in these resource to expand WordNet with new concepts. Accordingly, the task provides a high-quality dataset of one thousand definitions from a wide range of domains to be added to the WordNet hierarchy, either by adding them as new concepts or integrating them as new lemmas of an existing concept. The task provides a robust evaluation framework for measuring the accuracy of ontology expansion techniques. More broadly, the techniques developed as a part of Task 14 can play an important role in the construction of new automatically-built ontologies.

2 Task Description

The goal of Task 14 is to evaluate systems that enrich semantic taxonomies with new word senses drawn from other lexicographic resources. The task provides systems with a set of word senses that are not defined in WordNet.[1] Each word sense comprises three parts: a lemma, part of speech tag, and definition. For example, the noun *geoscience* is a word sense in our dataset which is associated with the definition "Any of several sciences that deal with the Earth". The word sense is drawn from Wiktionary.[2] For each of these word senses, a system's task is to identify a point in the WordNet's subsumption (i.e., is-a) hierarchy which is the most plausible point for placing the new word sense. In other words, a system's task is to find the most semantically similar WordNet synset to the given new word sense.

Operations Once the target synset is identified, a system has to decide how to integrate the new word sense. For a given new word sense s and a target synset $\mathcal{S}$ we define two possible operations:

- MERGE: when s refers to the same concept that is conceptualized by the synset $\mathcal{S}$. As a result of this operation s is added to the set of synonymous word senses in $\mathcal{S}$.

- ATTACH: when s refers to a more specific concept than $\mathcal{S}$. In other words, $\mathcal{S}$ is a generalization of the new word sense s (i.e., its hypernym). This operation creates a new synset containing the sole word sense s and attaches the new synset as a hyponym of $\mathcal{S}$ in the WordNet's subsumption hierarchy.

Table 1 shows example new word senses together with the target synset and the operation. Note that after both these operations, the polysemy of the lemma of s is increased by one. Also, the total number of synsets in the enriched WordNet increases by one after an ATTACH operation whereas it remains unchanged after MERGE, since in the latter case, a new word sense is added to an existing synset. Our datasets contain instances from noun and verb parts of speech.

2.1 Subtasks

For each item in our datasets, we provide the source dictionary from which the corresponding word sense (i.e., a word and its definition) is obtained. The participating systems were allowed to use the source dictionary in order to draw additional information or exploit its structural properties. Based on their usage of the source dictionary, we classify the participating systems into two categories:

- Resource-aware: the participating systems could use the URLs provided in the dataset to gather additional information (e.g., hyperlinks, wiki-markup) for performing the integration and may use additional information from any dictionary, including the one from which the target word sense had been obtained, e.g., Wiktionary.

- Constrained: the system might use any resource other than dictionaries.

We allowed each team to submit up to three runs per system type to let them explore different configurations, features, or parameter settings in the official rankings.

2.2 Related Tasks

Task 14 directly relates to three branches of prior tasks in SemEval. First, two recent tasks have evaluated automatic methods for constructing taxonomies (Bordea et al., 2015; Bordea et al., 2016). In these tasks, participants are presented with word pairs –but no glosses– and tasked with organizing the words into hypernym relationships. Task 14 provides the next step in such evaluations by explicitly

[1] We use WordNet 3.0.

[2] http://www.wiktionary.org

Lemma	POS	Definition	Target synset	Operation
geoscience	noun	Any of several sciences that deal with the Earth	earth_science – (any of the sciences that deal with the earth or its parts)	MERGE
mudslide	noun	A mixed drink consisting of vodka, Kahlua and Bailey's.	cocktail – a short mixed drink	ATTACH
euthanize	verb	To submit (a person or animal) to euthanasia.	destroy, put down – put (an animal) to death	MERGE
changing_room	noun	A room, especially in a gym, designed for people to change their clothes.	dressing_room – a room in which you can change clothes	MERGE
Apple	noun	An American multinational technology company headquartered in Cupertino, California, that designs, develops, and sells consumer electronics, computer software, online services, and personal computers.	corporation, corp – (a business firm whose articles of incorporation have been approved in some state)	ATTACH
own	verb	To illicitly obtain "super-user" or "root" access into a computer system thereby having access to all of the user files on that system.	crack – gain unauthorized access computers with malicious intentions.	ATTACH

Table 1: Sample instances from Task 14's datasets. A system's task is to identify, for a new word sense, the target synset and the corresponding operation.

incorporating polysemy into the task by requiring systems to specify a concept, rather than a word, as a hypernym. For example, when recognizing the relationships that a dog is a canine, the system would be required to specify that the concept should be attached to the animal sense of canine, not the tooth sense.

Second, the task of comparing a gloss associated with a new concept is closely related to the recent tasks on semantic similarity, i.e., Semantic Textual Similarity (Agirre et al., 2012; Agirre et al., 2013, STS) and Cross-Level Semantic Similarity (Jurgens et al., 2014, CLSS). Indeed, prior STS tasks included gloss pairs from OntoNotes in the datasets (Hovy et al., 2006) and CLSS had, among its four different evaluation types, an evaluation for systems measuring the similarity between word senses and words. However, while textual similarity is likely to be core component of Task 14 systems, the data is often richer than raw text by containing (a) regular linguistic structure where the parent concept is likely to be introduced first and (b) contextual features from where the gloss appears such as hyperlinks or example usages, which may help to disambiguate.

Third, prior tasks on Word Sense Induction (WSI) have evaluated methods that automatically discover the different meanings of a word (Manandhar et al., 2010; Jurgens and Klapaftis, 2013; Navigli and Vannella, 2013). However, the new senses discovered by these methods were never integrated into any taxonomy, making them difficult to use and relate to existing concepts. Task 14 provides a natural next step for WSI pairs, should any novel induced senses be matched with a gloss describing it.

3 Task Data

Given that WordNet 3.0 offers wide coverage of common concepts, the majority of novel concepts to be integrated are likely to come from topical domains, informal expressions, and neologisms. Therefore, the dataset for Task 14 was constructed to contain concepts from a wide variety of domains and to include glosses typical of those seen if performing an automated integration from online sources, such as those from heavily-curated sources such as Wiktionary and more idiosyncratic online glossaries with a single author. Table 3 shows the distribution of instances in the Task 14's training and test datasets across different genres. The dataset consists of a total of 1000 items, split into training and test datasets containing 400 and 600 items, respectively.

	Training			Test		
	M	A	Total	M	A	Total
Noun	27	322	349	26	490	516
Verb	6	45	51	6	78	84
Total	33	367	400	32	568	600

Table 2: The distribution of items in the task's datasets according to the part of speech and the target operation, i.e., Merge (M) and Attach (A).

Novel concepts were limited to nouns and verbs, as only these parts of speech have fully-developed taxonomies in WordNet.[3] Table 2 shows the distribution of training and test items according to their parts of speech and the intended operation, highlighting the fact that most new items are novel concepts that require a new synset to be added, rather than new lemmas to be included in an existing synset.

For each item, in addition to the target synset and the operation, we also provide the resource from which the new word sense was obtained. Glosses were provided as purely text data, with the hope was that systems may use the source URL provided with each gloss to identify additional page structure that could prove useful for concept integration (e.g., hyperlinks, wiki-markup, page topics).

3.1 Annotation Process

The two authors independently annotated each of the 1000 items, identifying the appropriate synset and operation. In a small number of cases, neither author could determine an appropriate integration for an item; such items were discarded and replaced with more-easily annotated items. Ultimately, all disagreements were discussed and adjudicated to determine the final dataset.

Annotators initially agreed on the annotation for 37.5% of the items. While this rate seems low at first glance, most disagreements were due to one annotator finding a more refined integration of the item, e.g,. DNA vs. Mutant Gene, which is expected given the large search space of over 82K noun and 13.7K verb synsets from which to find the appropriate hypernym or synonym synset. Indeed, most disagreements were very close in meaning; in fact, disagreements had an average semantic similarity between their synsets of 0.74 according to the Wu and Palmer (1994) measure. Hence, the moderate exact-match agreement is an underestimate of the true semantic agreement between annotators. Furthermore, several of the remaining dissimilar pairs were instances where similar concepts were distantly located in WordNet's structure.

The annotation proved difficult for three categories of concepts, not all of which were successfully integrated. First, many technical domains include unique processes and techniques specific to their field, e.g.,

> Lautering (noun) – The process of separating the sweet wort (pre-boil) from the spent grains in a lauter tun or with other straining apparatus.

However, some techniques and processes do not have a correspondence to any existing synsets, leaving their closest appropriate hypernym as a sense of process or technique. This difficulty is reflected in the current structure of WordNet, where *process#n#1* already has the dissimilar concepts of "fingerprinting," "computation," and "modus operandi" all as direct hyponyms, highlighting the challenge of placing some concepts. Where possible, we opted to avoid attaching new concepts to general senses, either by finding a more specific concept or leaving them out of the dataset entirely.

Second, in rare occasions, WordNet does not contain an intermediate concept necessary for the appropriate integration. For example, integrating the new concept

> Root (noun) – The administrative account (UID 0) on a *nix system that has all privileges; cf. superuser.

requires first having a concept of a computer account, which is not currently present in WordNet. These gaps are particularly evident for action nouns, where most verbs do not have a corresponding noun gerund. While the novel concept may still be attached to a more-distant hypernym of the appropriate location, this situation points to the need for an iterative integration process where intermediary concepts are first inserted.

[3] We do note that Tsvetkov et al. (2014) have proposed a taxonomy for adjectives, for which our methodology could be applied.

Third, concepts that express a negated or partial state often do not have associated concepts in the more-specific depths of the taxonomy. For example, annotators had difficulty finding appropriate synsets that were not too general for the following two concepts:

> NaN (noun) – Not a number; applied to numeric values that represent an undefined or unrepresentable value, such as zero divided by itself

> Neomort (noun) – A brain-dead human being that could be kept on life support for organ transplantation, medical and nursing education, and drug research.

Without additional synsets for representing partial or negated state, these concepts would need to be attached to very general synsets such as *value#n#1* or *person#n#1*.

The challenge of agreeing upon a specific location for a new concept underscores the need for automated approaches developed as a part of this task. As an ontology grows in size, it becomes less obvious where a new concept could be integrated, despite an annotator's familiarity with the concepts contained therein. Our annotation process relied on two annotators whose collective experience was necessary to identify the appropriate location. However, for larger ontologies such as BabelNet (Navigli and Ponzetto, 2012), which contains several orders of magnitude more concepts than WordNet, automated integration approaches will be necessary as it is infeasible for a single human annotator to recall the appropriate insertion point among millions of concepts.

3.2 Evaluation Metrics

We evaluated the performance of the participating system according to two criteria: (1) the accuracy by which the placements were performed, and (2) the percentage of items for which a decision was made (Recall).

3.2.1 Accuracy (Wu&P)

Our first criteria verifies the ability of a system to correctly identify the attachment or merge point in the WordNet hierarchy. Checking for exact matches would penalize equally both a placement in the near proximity of the intended synset and a random placement far in the network. A system's automatically-made attachment to the WordNet hierarchy is expected to be as close as possible to the correct attachment point given by the gold-standard data. We therefore evaluate the systems according to a fuzzy measure of accuracy which is sensitive to the distance between the intended target synset and the one outputted by the system. However, we recognize that links in the taxonomy do not necessarily represent uniform semantic distances, since siblings that are deep in the hierarchy tend to be more related to one another. Hence, a direct edge-counting approach might not provide a reliable basis for the evaluation of the attachment accuracy. Interestingly, the attachment accuracy evaluation can be cast as a WordNet-based semantic similarity measurement in which the goal is to compute the similarity between two concepts based on the structural properties of WordNet (Budanitsky and Hirst, 2006b), most important of which is the distance between the two. Therefore, we measure accuracy using the Wu and Palmer (1994, Wu&P) semantic similarity measure, defined as:

$$\frac{2 \cdot depth_{LCS}}{depth_1 + depth_2} \quad (1)$$

where $depth_1$ and $depth_2$ are the depths of the two concepts in WordNet's subsumption hierarchy (hypernymy/hyponymy relations) and $Depth_{LCS}$ is the depth of their least common subsumer, i.e., the most specific concept which is an ancestor of both the concepts. For each instance in the test set for which the system made a prediction, we measure the Wu&P similarity of the output attachment and the corresponding correct synset. An accurate system is expected to have a high similarity score when aggregated over all instances in the test set. Please note that picking the correct target synset with an incorrect operation is analogous to increasing the distance by one edge.

3.2.2 Lemma Match

A key challenge in the integration task is identifying the appropriate word in the gloss that denotes the hypernym (if it exists) and then disambiguating which sense of that word is the appropriate concept for attachment or merger. For example, given the

Genre	Subgenre	Training	Testing	Subgenre Total	Genre Total
Medical	UMLS	7	-	7	200
	Genomics	69	-	69	
	Virology	10	-	10	
	Dental	8	-	8	
	Healthcare	6	8	14	
	Immunology	-	24	24	
	Physiology	-	12	12	
	Homeopathy	-	6	6	
	Toxicology	-	3	3	
	Surgery	-	32	32	
	Veterinary medicine	-	4	4	
	Ophthalmology	-	5	5	
	Embryology	-	6	6	
Technical Language	Linux Glossary	20	5	25	200
	Mathematics	20	5	25	
	Narratology	10	14	24	
	Earth Science	10	16	26	
	Music	-	41	41	
	Brewing	-	35	35	
	Neuroscience	-	14	14	
	Architecture	-	10	10	
Sports domains	Gridiron football	25	-	25	100
	Cycling	25	-	25	
	Golf	-	13	13	
	Sailing	-	8	8	
	Weightlifting	-	5	5	
	Climbing	-	9	9	
	Volleyball	-	15	15	
Legal Language	American Law	-	100	100	100
Slang	American Slang	30	35	60	150
	British Slang	5	5	10	
	Online Slang/jargon	25	15	40	
	Military Slang	10	10	20	
Jargon	Computer	20	50	50	50
Idioms	American Idioms	20	25	50	50
Religious Language	Islam	15	10	25	100
	Hinduism	15	10	25	
	Judaism	20	5	25	
	Catholicism	15	10	25	
Financial Language	Banking	15	10	25	50
	Stocks	-	25	25	
Total		400	600	1000	1000

Table 3: The distribution of instances across different genres in the training and test data sets of Task 14.

item

> Grief (verb) – To deliberately harass and annoy or cause grief to other players of a game in order to interfere with their enjoyment of it

a system may correctly identify that the verb *harass* is the hypernym but select the wrong sense to which *grief* should be attached. Such a mistake would be penalized heavily according to the Wu&P measure and mask that the system is accurate at identifying hypernyms in glosses. Therefore, we include a second unofficial metric, lemma match, that measures the percentage of items for which the system has selected a synset with at least one word in common with the correct synset where the item should be integrated; i.e., how often the system picked the right word but wrong sense.

3.2.3 Recall

Some word senses may be more difficult to place in the WordNet hierarchy than others due to a variety of reasons, such an entry with a gloss that contains many out of vocabulary words. Therefore, we allow a system to decline to place these senses in order to avoid making placements with low confidence. As an evaluation metric, we report Recall as the percentage of items for which a decision was made by the system.

3.3 System ranking

A system's performance is computed by the F1 score of Wu&P and Recall. The official ranking of the systems was done according to their F1 scores.

4 Systems

Five teams submitted 13 systems, where each team's systems were variations on a common architecture. No system utilized resource-specific features beyond the gloss (e.g., the Wiktionary markup) and so all systems were ultimately submitted in the constrained category. Systems were compared against two baselines.

4.1 Participants

The **MSejrKU** systems build definitional representations based on skip-gram vectors trained on Wikipedia data and incorporates syntactic features. Words in a candidate gloss are disambiguated using the method of Agirre and Soroa (2009) and then a classifier predicts the goodness of fit for a candidate attachment synset related to those in the gloss.

The **Duluth** systems perform string matching to compare a definition with each of the glosses in WordNet. Given a new definition, systems differ in which words are included from the WordNet synset for comparison: Duluth2 uses only the words in the definition after stopword removal, while Duluth1 extends Duluth2 by including words from the hypernyms of the compared synset. Duluth3 extends Duluth1 with words from the hyponyms but also takes the step of breaking each definition into character tri-grams to capture surface-form regularities. The **UMNDuluth** team performs a similar approach but weights gloss similarity by favoring specific kinds of terms, such as those that are longer and those that appear in WordNet.

The **TALN** systems project the definition of the novel term into a vector space using SENSEMBED (Iacobacci et al., 2015). Then this vector is compared with the vectors for senses in WordNet to find the closest match. System variations address issues when words have no associated vectors and how to select between candidate attachments.

The **JRC** system uses a form of second-order similarity by representing each definition as a vector over the synsets that contain its words. New terms are attached by finding the WordNet synset whose definition has maximal cosine similarity.

The **VCU** systems adopt multiple approaches based on textual similarity. Run1 uses a second-order expansion by representing a definition using frequency of words related to those in the definition. Run2 compares glosses using Lesk relatedness measure. Run3 performs no pre-processing and compares the words in the glosses directly as first-order vectors.

4.2 Baselines

The first baseline, **Random synset** captures the expected performance of a system at chance when attaching the new concept to a randomly picked synset from WordNet with the appropriate part of speech. This baseline provides the lower bound in expected similarity for an attachment.

The second baseline captures our observation that

Rank	Team	System	LM	Wu&P	Recall	F1
1	MSejrKU	System2	0.428	0.523	0.973	0.680
2	MSejrKU	System1	0.432	0.518	0.968	0.675
3	TALN	test_cfgRun1	0.360	0.476	1.000	0.645
4	TALN	test_cfgRunPickerHypos	0.240	0.472	1.000	0.641
5	TALN	test_cfgRun2	0.353	0.464	1.000	0.634
6	VCU	Run3	0.161	0.432	0.997	0.602
7	VCU	Run2	0.171	0.419	0.997	0.590
8	VCU	Run1	0.124	0.408	0.997	0.579
9	Duluth	Duluth2	0.043	0.347	1.000	0.515
10	JRC	MainRun	0.066	0.347	0.987	0.513
11	Duluth	Duluth3	0.017	0.345	1.000	0.513
12	UMNDuluth	Run1	0.098	0.340	0.998	0.507
13	Duluth	Duluth1	0.023	0.331	1.000	0.498
Baseline: First word, first sense			0.415	0.514	1.000	0.679
Baseline: Random synset			0.000	0.227	1.000	0.370

Table 4: Evaluation results showing the Lemma Match (LM), Wu&P, and Recall measures.

glosses are reasonably well structured such that the word expressing the hypernym concept appears early in the gloss (if at all). Therefore, given new word sense s with definition d_s and part of speech tag p, the **First word, first sense** (FWFS) baseline picks the first occurring word w in d_s with part of speech p as the hypernym (i.e., the first noun if the word sense to be attached is a noun and the first verb otherwise).

The new word sense is then attached to the synset containing the first sense of w. For example, given the item

> Immunoglobin (noun) – Any protein that functions as an antibody

the FWFS baseline attaches the item to the first sense of the noun protein in WordNet, i.e., *protein#n#1*. Despite the wide variety of domains seen in the data, 65% of all integrations in the gold standard data connect the first sense of the target word, suggesting that in the absence of specific information to disambiguate a word in the gloss, its first (most frequent) sense is relatively high precision back-off strategy.

For the FWFS baseline, glosses are POS-tagged using CoreNLP (Manning et al., 2014) and we include a minimal heuristic that prevents attaching to "a" or "an," both of which are nouns in WordNet. In the rare event that no word can be found with the same part of speech, the item is attached to either the general concepts of *entity#n#1* or *be#v#1*, depending on the part of speech.

5 Results

All of the thirteen participating systems improved over the Random synset baseline. Table 4 shows the evaluation results for Task 14's participating systems. However, only one of the systems, System2 of MSejrKU, could slightly outperform the FWFS baseline, showing the competitiveness of this simple baseline, which takes advantage of the inherent structure of definitions. Indeed, the lower ranked systems frequently performed holistic comparisons between gloss texts, which frequently include terms from later in the gloss that do not aid in identifying the closest meaning.

While its performance is relatively high among participants, the FWFS baseline should not be mistaken for a satisfactory solution; many of the attachments made by the baseline are overly general and do not take advantage of the remainder of the gloss's content, which can identify the correct, more-specific concept to which the item should be attached. For example, with the item

> Hot reactor (noun) – A person whose blood pressure and heart rate increase abnormally in response to stress

the baseline naively attaches to *person#n#1*, while a

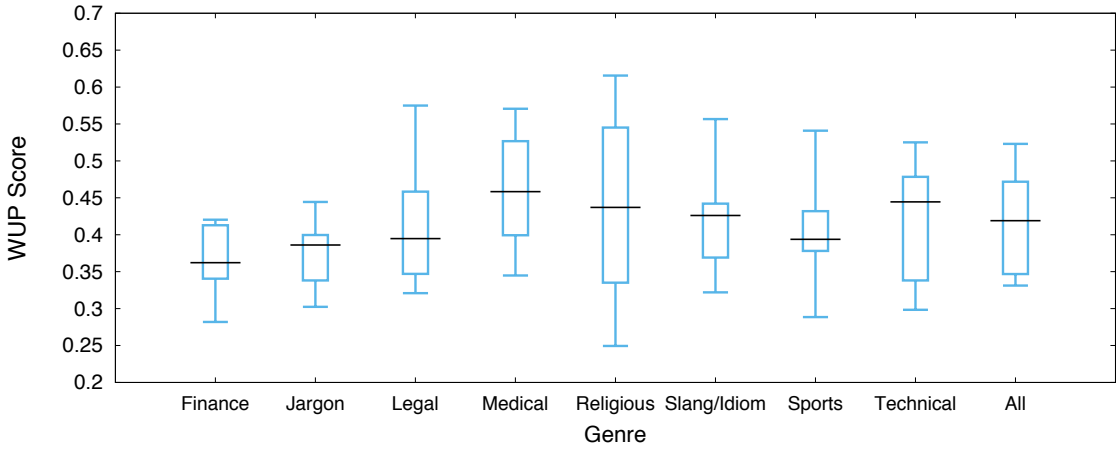

Figure 1: The distribution of Wu&P scores of the participating systems per genre. Whiskers show minimum and maximum scores and lines denote the median.

more sophisticated solution could use the additional text in gloss to identify an appropriate hyponym of *person#n#1* to which the item may be attached, e.g., *sick_person#n#1*. Thus, we speculate that the gloss similarity used by participants may still prove highly useful by first identifying the appropriate general concept in the gloss (e.g., as the baseline does) and then searching its hyponyms for a better match.

Examining the performances of systems in Table 4, we see that no system performed significantly better on the Lemma Match measure than Wu&P, with both measures being highly correlated at r=0.96. This suggests that when the appropriate hypernym lemma was present in a gloss, systems struggled most with selecting it as the correct candidate lemma in the gloss, rather than identifying which synset of that lemma was the correct attachment.

Given the variety of genres and sources from which new definitions were drawn, we performed a follow-up analysis to examine the impact of the genre on system performance. Figure 1 shows the distribution of scores per genre. Surprisingly, Religious definitions cause the most variance among systems and also saw the highest and lowest system scores per genre. Religious definitions were drawn from more sources beyond just Wiktionary and thus, such variance may reflect systems' robustness to different writing styles. Systems performed worst on the Finance and Jargon domains; however, both genres had little training data relative to testing data, suggesting that systems had difficulty generalizing from few examples. Nevertheless, systems still performed well for the Legal genre which was held out as a surprise dataset with no training data.

6 Conclusion

Semantic taxonomies are core components of many NLP systems and multiple approaches have been proposed for how to extend such taxonomies automatically with new concepts. We have introduced SemEval-2016 Task 14 as a framework and dataset for evaluating the accuracy of systems at integrating new definitions as concepts into an ontology using WordNet 3.0 as a base resource. Five teams submitted 13 systems for participation, with all teams performing better than chance but only one team surpassing a simple baseline that leverages knowledge of the expected word order in a definition to guess the correct hypernym concept. Our results point towards significant opportunity for improving taxonomy enrichment. In future work, we intend to integrate the best insights of this task into the next version of CROWN,[4] an automatically constructed extension of WordNet with concepts from online glossaries and lexicographic resources.

References

Eneko Agirre and Aitor Soroa. 2009. Personalizing pagerank for word sense disambiguation. In *Proceed-*

[4] https://github.com/davidjurgens/crown

ings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, pages 33–41. Association for Computational Linguistics.

Eneko Agirre, Mona Diab, Daniel Cer, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 task 6: A pilot on semantic textual similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 385–393. Association for Computational Linguistics.

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. sem 2013 shared task: Semantic textual similarity, including a pilot on typed-similarity. In *In* SEM 2013: The Second Joint Conference on Lexical and Computational Semantics. Association for Computational Linguistics*. Association for Computational Linguistics.

Georgeta Bordea, Paul Buitelaar, Stefano Faralli, and Roberto Navigli. 2015. SemEval-2015 task 17: Taxonomy extraction evaluation (texeval). In *Proceedings of the 9th International Workshop on Semantic Evaluation*. Association for Computational Linguistics.

Georgeta Bordea, Els Lefever, and Paul Buitelaar. 2016. SemEval-2016 task 13: Taxonomy extraction evaluation (texeval-2). In *Proceedings of the 10th International Workshop on Semantic Evaluation*. Association for Computational Linguistics.

Alexander Budanitsky and Graeme Hirst. 2006a. Evaluating WordNet-based measures of Lexical Semantic Relatedness. *Computational Linguistics*, 32(1):13–47.

Alexander Budanitsky and Graeme Hirst. 2006b. Evaluating WordNet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47.

Andrea Esuli and Fabrizio Sebastiani. 2006. SENTIWORDNET: A publicly available lexical resource for opinion mining. In *In Proceedings of the 5th Conference on Language Resources and Evaluation*, pages 417–422.

Hui Fang. 2008. A re-examination of query expansion using lexical resources. In *Proceedings of ACL-08: HLT*, pages 139–147, Columbus, Ohio.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Database*. MIT Press, Cambridge, MA.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: the 90% solution. In *Proceedings of the human language technology conference of the NAACL, Companion Volume: Short Papers*, pages 57–60. Association for Computational Linguistics.

Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. 2015. SensEmbed: learning sense embeddings for word and relational similarity. In *Proceedings of ACL*, pages 95–105.

David Jurgens and Ioannis Klapaftis. 2013. SemEval-2013 task 13: Word sense induction for graded and non-graded senses. In *Second joint conference on lexical and computational semantics (* SEM)*, volume 2, pages 290–299.

David Jurgens and Mohammad Taher Pilehvar. 2015. Reserating the awesometastic: An automatic extension of the WordNet taxonomy for novel terms. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1459–1465, Denver, Colorado.

David Jurgens, Mohammad Taher Pilehvar, and Roberto Navigli. 2014. SemEval-2014 task 3: Cross-level semantic similarity. In *Proceedings of the* 8^{th} *International Workshop on Semantic Evaluation (SemEval 2014), in conjunction with COLING 2014*, pages 17–26, Dublin, Ireland.

Jey Han Lau, Paul Cook, Diana McCarthy, David Newman, and Timothy Baldwin. 2012. Word sense induction for novel sense detection. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 591–601. Association for Computational Linguistics.

Suresh Manandhar, Ioannis P Klapaftis, Dmitriy Dligach, and Sameer S Pradhan. 2010. SemEval-2010 task 14: Word sense induction & disambiguation. In *Proceedings of the 5th international workshop on semantic evaluation*, pages 63–68. Association for Computational Linguistics.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *ACL (System Demonstrations)*, pages 55–60.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Roberto Navigli and Daniele Vannella. 2013. SemEval-2013 task 11: Word sense induction and disambiguation within an end-user application. In *Second Joint Conference on Lexical and Computational Semantics (* SEM)*, volume 2, pages 193–201.

Roberto Navigli. 2009. Word Sense Disambiguation: A survey. *ACM Computing Surveys*, 41(2):1–69.

Mohammad Taher Pilehvar, David Jurgens, and Roberto Navigli. 2013. Align, Disambiguate and Walk: a Unified Approach for Measuring Semantic Similarity. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1341–1351, Sofia, Bulgaria.

Michael Poprat, Elena Beisswanger, and Udo Hahn. 2008. Building a BioWordNet by using WordNet's data formats and WordNet's software infrastructure: a failure story. In *Proceedings of the Workshop on Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 31–39, Columbus, Ohio.

Rion Snow, Dan Jurafsky, and Andrew Ng. 2006. Semantic taxonomy induction from heterogeneous evidence. pages 801–808.

Antonio Toral, Rafael Muoz, and Monica Monachini. 2008. Named Entity WordNet. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC)*, pages 741–747.

Yulia Tsvetkov, Nathan Schneider, Dirk Hovy, Archna Bhatia, Manaal Faruqui, and Chris Dyer. 2014. Augmenting English adjective senses with supersenses. In *LREC*. European Language Resources Association.

Giannis Varelas, Epimenidis Voutsakis, Paraskevi Raftopoulou, Euripides G.M. Petrakis, and Evangelos E. Milios. 2005. Semantic similarity methods in WordNet and their application to information retrieval on the web. In *Proceedings of the 7th Annual ACM International Workshop on Web Information and Data Management*, pages 10–16.

Zhibiao Wu and Martha Palmer. 1994. Verbs semantics and lexical selection. In *Proceedings of the 32Nd Annual Meeting on Association for Computational Linguistics*, pages 133–138.

Ichiro Yamada, Jong-Hoon Oh, Chikara Hashimoto, Kentaro Torisawa, Junichi Kazama, Stijn De Saeger, and Takuya Kawada. 2011. Extending WordNet with hypernyms and siblings acquired from wikipedia. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*, pages 874–882.