Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 3121–3127
Marseille, 11–16 May 2020

Inducing Universal Semantic Tag Vectors

Da Huo, Gerard de Melo
Department of Computer Science, Rutgers University–New Brunswick, NJ, USA
dh637@scarletmail.rutgers.edu, gdm@demelo.org

Abstract

Given the well-established usefulness of part-of-speech tag annotations in many syntactically oriented downstream NLP tasks, the recently proposed notion of semantic tagging (Bjerva et al., 2016) aims at tagging words with tags informed by semantic distinctions, which are likely to be useful across a range of semantic tasks. To this end, their annotation scheme distinguishes, for instance, privative attributes from subsective ones. While annotated corpora exist, their size is limited and thus many words are out-of-vocabulary words. In this paper, we study to what extent we can automatically predict the tags associated with unseen words. We draw on large-scale word representation data to derive a large new Semantic Tag lexicon. Our experiments show that we can infer semantic tags for words with high accuracy both monolingually and cross-lingually.

Keywords: semantic tagging, word vectors, multilingual

1. Introduction

Lexical category distinctions have been studied since the beginnings of linguistics. Traditional part-of-speech tagging has focused on distinctions based on the grammatical function of words, i.e., the syntactic role that words play within a sentence.

Semantic Tags. In contrast, the recently proposed task of *Semantic Tagging* (Bjerva et al., 2016) considers a set of tags that are informed by *semantic* distinctions conjectured to be pertinent for semantic parsing and other semantics-driven tasks.

The annotation scheme distinguishes, for instance, privative attributes (`PRI`) such as *former*, *fake* from intersective ones (`IST`) such as *vegetarian*, and subsective ones (`SST`) such as *skillful*, making it easier for a system to discern that a *fake* detective is not a detective. Different kinds of named entities are distinguished (e.g., people, geopolitical entities, artifacts, etc.), as in named entity recognition, including temporal categories (e.g., there are separate tags for decades, day of week, etc.). Moreover, there are dedicated tags for different tenses of events (such as past tense *ate*), tense & aspect markers (e.g., *was* in *was reading*), roles (*victim*), implications (*if*, *unless*), greetings (*hi*, *bye*), and many others. At the same time, more syntactically motivated distinctions, such as between adjectives and adverbs, are disregarded in this annotation scheme. In follow-up work, Abzianidze and Bos (2017) presented an improved tag set and showed that the tags exhibit the potential to apply cross-lingually.

Motivation. In terms of available data, the Parallel Meaning Bank project (Abzianidze et al., 2017) provides a parallel corpus covering four languages (English, Dutch, German, and Italian) with rich annotations based on Discourse Representation Theory. At the lexical level, it includes semantic tag annotations, which we rely on in our work.

However, due to the novelty of this task, the available annotations are limited in quantity and consist of a mix of gold and silver standard data. Just up to a few thousand sentences per language have been manually annotated. Thus, the vocabulary coverage of this data is limited in the sense that, for many words, we do not know what tags may be relevant for them.

Goals. In this paper, we study to what extent simple automated methods can be invoked to infer the associated tags for previously unseen words. We aim at a lexical resource that reveals the semantic tags associated with a much larger set of words than given in the annotated corpus. The tag distribution for a word can be viewed as an interpretable vector representation.

Overview. We predict such interpretable representations by drawing on the annotated seed corpus in conjunction with information about word relatedness from larger-scale word vector representation data.

Our experiments show that our method infers semantic tags for unseen words with high accuracy for four languages. This can finally be used to induce a large new Semantic Tag lexicon, providing semantic tag vectors for millions of words and names. Additionally, we can induce vectors cross-lingually for numerous additional languages.

2. Related Work

Interpretable Lexical Representations. Brants (2000) highlighted the importance of handling unknown words in part-of-speech tagging. Our work has similar goals to the research by Cucerzan and Yarowsky (2000) on estimating part-of-speech probabilities for unseen words based on probabilities for known words.

Faruqui and Dyer (2015) consulted a range of lexical resources to create non-distributional vectors that capture numerous different properties of words. These vectors are thus fairly high-dimensional.

Recent work has considered lower-dimensional interpretable vectors focusing on particular aspects of words. For instance, Dong and de Melo (2018b) induced vectors that capture the sentiment polarity of words in different domains, **?**) developed interpretable vectors reflecting emotional associations of words, and Shoeb et al. (2019) proposed vectors reflecting associations with emojis.

In this paper, we seek to induce interpretable vectors reflecting semantic tag distributions.

Cross-Lingual Lexical Representations. A number of different cross-lingual projection approaches have been proposed, from simple linear projection matrices (Mikolov et al., 2013) to more powerful GAN-based techniques (Fu et al., 2020). Such methods assume that one has already obtained vector representations for two languages, which subsequently need to be aligned, or that one has a parallel corpus that can be exploited for joint training of bilingual embeddings covering two languages (de Melo, 2017b).
In our work, we instead adopt the use of a translation dictionary. Previous work has considered dictionary data from a massively multilingual graph perspective (de Melo, 2015; de Melo, 2017a). In this paper, we do not project dense word vector representations but instead seek interpretable tag distribution vectors, similar to the interpretable sentiment vectors considered by Dong and de Melo (2018a). Rather than invoking a graph-based approach with mutual interactions, we show that a single hop translation lookup approach is sufficient to obtain high-quality predictions.

3. Method

In the following, in Section 3.1, we first describe our procedure for inducing semantic tag vectors monolingually, for unseen words within the same language. Subsequently, we briefly describe an extension for cross-lingual prediction in Section 3.2.

3.1. Creating Tag Vectors

Our procedure to create semantic tag vectors for unseen words consists of a few simple steps. First, we consider an annotated corpus labelled with semantic tags to extract the distribution of tags for each word. Then we draw on large-scale word vectors to be able to propagate information from words observed in the corpus to new unseen words, based on nearest neighbour retrieval. The semantic tag distributions of suitable neighbours are aggregated to infer a semantic tag distribution for the new word.

3.1.1. Input

As input, we assume an annotated seed corpus $\mathcal{C} = ((w_1, t_1), \ldots, (w_L, t_L))$, in which word tokens w_i have been annotated with tags $t_i \in \mathcal{T}$ from a global tag set $\mathcal{T}$. Here, the vocabulary of tokens $\mathcal{V} = \bigcup_{w_i \in \mathcal{C}} \{w_i\}$ contains either just raw string tokens or (string,part-of-speech) tuples. The latter will allow us to make predictions that account for the part-of-speech (POS) of a word, such that we can also derive separate tag predictions for the noun *bear* as opposed to the verb *to bear*, for example.

3.1.2. Seed Tag Vectors

Given this seed data, we first compute seed tag vectors $\mathbf{v}_w$ for $w \in \mathcal{V}$ as follows:

$$\mathbf{v}_w = \sum_{(w_i, t_i) \in \mathcal{C}:\, w_i = w} \mathbf{e}(t_i) \tag{1}$$

Here, $\mathbf{e}(t)$ is a function yielding a $|\mathcal{T}|$-dimensional vector with a one-hot encoding of tag t. The resulting semantic tag vectors hence capture the distribution of a word's tags across the corpus. Each dimension of the vector space corresponds to a different semantic tag in $\mathcal{T}$, and a seed vector's entry in a given dimension reflects how often the word was encountered in the corpus as being labelled with the respective tag.

3.1.3. Neighbour Retrieval

Having computed such tag vector representations for each word observed in our annotated corpus $\mathcal{C}$, we now wish to infer similar kinds of semantic tag vectors $\mathbf{v}_w$ for unseen input words w.
For this, we draw on a preexisting word embedding matrix E providing regular word vector representations for words. This data is entirely separate from the annotated seed corpus. It will normally be derived from a corpus that is many orders of magnitude larger.
Given a target word w, we first determine its k nearest neighbours $N_k(w)$ using the preexisting word embeddings E as follows:

$$N_k(w) = \sigma_k(w, E, \mathcal{V}_E \cap \mathcal{V}) \tag{2}$$

Here, σ_k is a function yielding a set of the k closest words in the embedding space E with respect to a vocabulary. In particular, it first retrieves the regular word vector $\mathbf{u}_w$ for the input word w in the word embedding space E, and computes its k nearest neighbours in E in terms of the Euclidean distance. However, for this, σ_k only considers the subset of words that are in $\mathcal{V}_E \cap \mathcal{V}$, i.e., those words that are not only in the vocabulary $\mathcal{V}_E$ of E, but simultaneously also in the annotated corpus vocabulary $\mathcal{V}$. All other words are disregarded, as we do not have tag information for them. We study 2 different instantiations of σ_k.

Form-based Prediction. We consider an ordinary word embedding table E that provides vectors for word forms, without specifically accounting for the part-of-speech properties of words.

POS-aware Prediction. We invoke word embeddings E trained to provide embeddings for specific lemma and part-of-speech combinations. In this approach, σ_k retains only those vocabulary items with a compatible part-of-speech tag to w, while still seeking to achieve that $|N_k(w)| = k$ to the extent possible.

3.1.4. Tag Vector Induction

Finally, based on the neighbours, we predict semantic tag vectors $\mathbf{v}_w$ for unseen words w as

$$\mathbf{v}_w = \frac{1}{|N_k(w)|} \sum_{w' \in N_k(w)} \frac{\mathbf{u}_w\, \mathbf{u}_{w'}}{\|\mathbf{u}_w\|\, \|\mathbf{u}_{w'}\|}\, \mathbf{v}_{w'}. \tag{3}$$

Here, the various $\mathbf{v}_{w'}$ are semantic tag vectors for the k nearest neighbours in N_k, and their contribution is weighted based on the cosine similarity of the corresponding vectors $\mathbf{u}_w$ as given by the preexisting embedding matrix E. Importantly, $\mathbf{v}_{w'}$ are not normalized, and thus w' observed more frequently in the seed data can contribute to the prediction to a greater extent. Optionally, one may normalize the predicted vectors such that they reflect a predicted probability distribution over $\mathcal{T}$.

3.2. Cross-Lingual Induction

Our approach can also be extended for cross-lingual semantic tag vector induction. We use the same vector computation as above but need to take one additional step. Given

a non-English target word w_0, we first retrieve a set of English translations

$$W = \{w \mid (w_0, w) \in T, w \in \mathcal{V}_E\} \quad (4)$$

from a translation dictionary T, which we assume provides part-of-speech specific translations. In other words, during this process, we only consider translations w that occur in the vocabulary of our English embeddings E with a matching part-of-speech tag. If $|W| > 0$, we can then obtain

$$\mathbf{v}_{w_0} = \frac{1}{|W|} \sum_{w \in W} \mathbf{v}_w, \quad (5)$$

where the $\mathbf{v}_w$ are tag vectors available as seed vectors or computed using the monolingual method from Eq. 3.

4. Experiments

The merits of the above approach are evaluated in a series of English language and cross-lingual experiments, including detailed analyses for specific categories of words.

4.1. Data

Seed Corpus. We rely on version 2.1.0 of the Parallel Meaning Bank (PMB) corpus (Abzianidze et al., 2017), of which we consider the gold quality subsets, which are the parts of the data that were fully verified by human annotators. Statistics of this data are given in Table 1. The corpus provides data for four languages: English, German, Italian, and Dutch. Note that due to updates to the tagging scheme, the tag inventory in the corpus differs in certain minor ways from the definitions provided in the original papers.

Preexisting Embedding Matrix E. In terms of word vectors, we rely on the standard Stanford GloVe 300-dimensional word vectors (Pennington et al., 2014) as the embeddings E for the Form-based Prediction approach from Section 3.1.3. We use Sketch Engine English (Web, 2013, 20 billion tokens) Lemma + Part of Speech word embeddings[1] as E for the POS-aware Prediction approach. The latter has a vocabulary size of 6,143,073.

Translations. As for our cross-lingual evaluations, we rely on translations extracted from the 2018-11-20 English edition of Wiktionary using a custom extraction framework (de Melo, 2014). This data allows us to obtain English translations of non-English input words, for which we search both possible translation directions.

4.2. Experimental Protocol

We tested our method on English as well as cross-lingually using gold data from the PMB corpus. For English words, besides the 90-dimensional semantic tag vector prediction experiments, we also conducted experiments to predict more general 15-dimensional coarse-grained tags for the corpus (denoted as *English (C)*). These coarse-grained tags, listed in Table 4, are provided along with the Universal Semantic Tagging data as a high-level categorization of tags.
Due to the small size of the seed data, we rely on a leave-one-out evaluation for each setting, i.e., we consider the gold data as the ground truth and try to predict each word's ground truth tag vector separately, based only on other seed words in the data, excluding the target word itself. In the cross-lingual case, we predict tag vectors for non-English using English tag vectors, but consider only non-English tokens for which we have suitable English translations for which we can obtain or predict such tag vectors.
We rely on the *average cosine similarity* score between the ground truth vector and the predicted vector to quantify the accuracy of our method.

[1] https://embeddings.sketchengine.co.uk/

4.3. Overall Results

4.3.1. Results for Form-based Prediction

For our Form-based Prediction approach, the experimental results are given in Table 2. Our approach obtains reasonably high cosine similarities, but they are better for *English (C)*, i.e., at the coarse-grained level. The results are also reasonably strong on cross-lingual mappings, despite the fact that this involves relying on a translation dictionary, which may bring in additional ambiguity and may result in a skewed tag distribution if the set of available translations is skewed towards particular word senses.
Figure 1 presents the average cosine similarity score for different k visually to make the trends more obvious. Recall that k is the number of nearest neighbours used for σ_k when predicting semantic tag vectors. Initially, increasing k helps for more robustness, but particularly large k may lead to the inclusion of semantically distant neighbours.

4.3.2. Results for POS-aware Prediction

For the POS-aware Prediction, the experimental results are given in Table 3. Here, we find substantially improved results over the Form-based approach. In fact, the fine-grained English prediction with 90 tags becomes about as accurate as the coarse-grained prediction. This suggests that part-of-speech tags, despite their syntactic nature, aid in discriminating between semantically ambiguous forms.
As plotted in Figure 2, we tend to observe improved scores as k grows. Increasing k leads to more robust results than when betting on just 1 or 2 neighbours, and the POS-based method eliminates some of the semantically remote candidates that the Form-based method might consider.
For our coarse-grained prediction in particular, we observe that the prediction accuracy continuously increases as k increases. We found that for most of the words in our test data, the coarse-grained prediction works well and often leads to an average cosine similarity of 1. For ambiguous words, such as *well*, the neighbours we get often vary in terms of their meaning. Therefore, more neighbours often lead to a higher probability that the neighbours' meaning match the input's meaning. For instance, the prediction accuracy for the word *well* increases from 0.0 to 0.45 as we increase k from 3 to 20.

4.4. Fine-Grained Analysis

4.4.1. Analysis per Coarse-Grained Semantic Tag

Table 4 reports separate averages for each coarse-grained semantic tag while predicting fine-grained English word tags using POS-aware Prediction with $k = 3$, We observe that our method works particularly well on entities, which

Language	Documents	Sentences	Tokens (Distinct)
English	4,555	4,567	2,7433 (4,039)
German	1,175	1,176	6,459 (1,823)
Italian	635	635	3,315 (1,075)
Dutch	586	587	3,354 (1,074)

Table 1: Parallel Meaning Bank Gold Data Statistics

Setup	Language	1	2	3	4	5	10	20
Monolingual	English	0.61	0.62	0.64	0.65	0.66	0.65	0.63
	English (C)	0.79	0.78	0.79	0.79	0.79	0.77	0.74
Cross-lingual	German	0.65	0.68	0.71	0.70	0.71	0.73	0.73
	Dutch	0.62	0.67	0.67	0.67	0.69	0.68	0.67
	Italian	0.62	0.65	0.66	0.67	0.67	0.68	0.68

Table 2: Form-based Prediction for different k

Setup	Language	1	2	3	4	5	10	20
Monolingual	English	0.76	0.77	0.78	0.78	0.78	0.77	0.78
	English (C)	0.75	0.76	0.76	0.76	0.77	0.77	0.78
Cross-Lingual	German	0.84	0.84	0.83	0.83	0.83	0.84	0.84
	Dutch	0.86	0.84	0.84	0.85	0.84	0.84	0.83
	Italian	0.82	0.84	0.82	0.83	0.83	0.86	0.86

Table 3: POS-aware Prediction results for different k

Tag	Description	Score	Count
NAM	named entities	0.77	497
UNE	unnamed entities	0.87	1,594
TIM	temporal entities	0.82	137
ATT	attribute	0.82	607
DSC	discourse	0.96	6
EVE	events	0.51	1,271
ANA	anaphoric	0.53	53
ACT	speech act	0.00	11
COM	comparative	0.31	33
DEM	demonstrative	0.38	8
LOG	logical	0.25	82
MOD	modality	0.40	82
TNS	tense	0.17	54
DXS	deixis	N/A	0
UNK	unknown	N/A	9

Table 4: Average Cosine Similarity Score per Coarse-Grained Semantic Tag in Gold Data

constitute the majority of PMB's gold data. These include unnamed entities (UNE), named entities (NAM), and temporal entities (TIM). Our method further attains a high accuracy on attributes (ATT), including colors, degrees, scores, and quantities, despite the fine-grained semantic distinctions mentioned in Section 1. It fares slightly worse on ANA (anaphoric) and EVE (events) such as untensed simple, present simple, and past simple ones. These might not be sufficiently well-distinguished in the regular word vectors.

POS	Score	Count
adverb	0.48	96
pronoun	0.65	36
preposition	0.85	46
adjective	0.85	404
noun	0.82	2,177
verb	0.58	1,237
other	0.57	237

Table 5: Average Cosine Similarity Score per Part-of-Speech Tag in Gold Data

4.4.2. Analysis by Part-of-Speech

We also assessed the quality for different part-of-speech tags of the ground truth English language fine-grained tag vectors based on the PMB gold data. The break-down of results for POS-aware Prediction with $k = 3$ is given in Table 5. We find that our prediction method achieves a high accuracy on nouns and adjectives, which make up 60% of our testing dataset. It performs slightly worse on verbs and other categories. We observed that this in part stems from the fact that the corpus often just has a single occurrence of a particular form of a verb, so the ground truth vectors for it do not reflect the overall distribution of possible tags for the word, but just a single observed tag. Hence, the scores underestimate how well our method predicts the actual distribution. The prediction is also poorer on a few other classes, which, however, are extremely infrequent in the corpus.

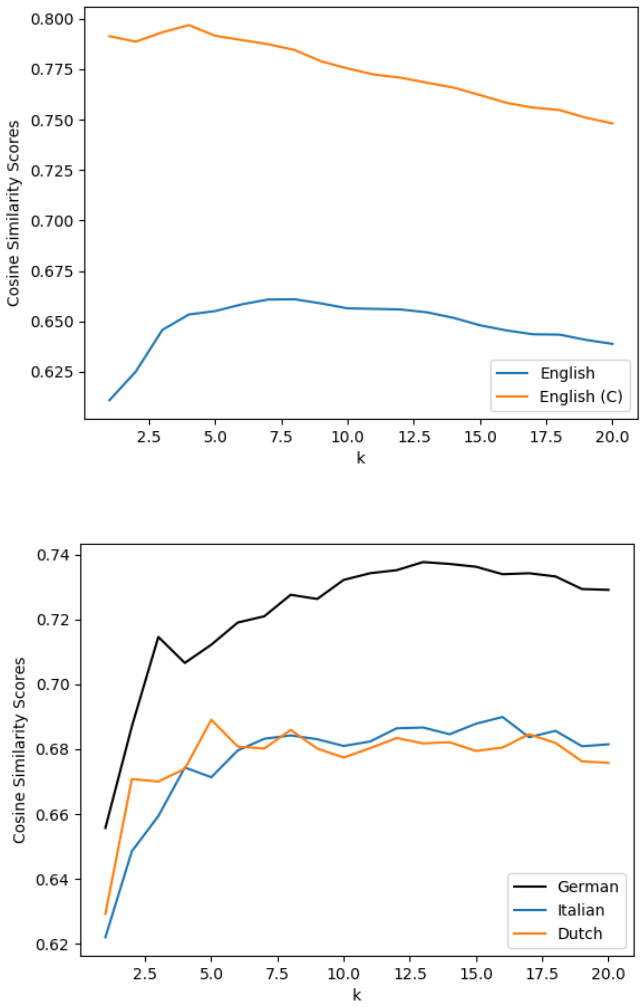

Figure 1: Cosine similarity score vs. k for Form-based Prediction (top: English, bottom: cross-lingual)

Language	**# Words**
Finnish	80,831
Russian	79,939
German	72,574
Mandarin Chinese	60,209
French	57,164
Spanish	55,721
Portuguese	51,219
Italian	46,999
Dutch	41,455
Japanese	37,664
...	...
All	1,523,736

Table 6: Coverage of Cross-Lingual Tag Vectors

4.4.3. Error Analysis

We observed that the principal reason for the poorer results of the Form-based approach in comparison with the POS-aware one is that the nearest neighbour structure of the GloVe vectors diverges from what is needed for semantic tag distinctions. The neighbours computed from GloVe vectors are the most similar words that could be put into a similar context while neglecting the word order in the context. Often, these are other forms of the same lemma.

For instance, the nearest neighbour of *notice* is *notices*. While these are closely related, semantic tagging makes fine-grained distinctions. The semantic tags assigned to these two words may diverge due to the ambiguous part-of-speech categories. Some occurrences of *notices*, for instance, may be tagged as `CON` (*concept*), while some occurrences of *notice* may be classified as `EXS` (*untensed simple*). Evidently, part-of-speech information can aid in resolving much of this sort of ambiguity between semantic tags. Hence, our exploration of POS-aware Prediction, which accounts for it.

For our English POS-aware Prediction (k=3), we further analyzed our results by generating a 90×90-dimensional confusion matrix. Among a total of 4,039 distinct words in the gold data, words tagged with `NOW` (*present tense*) are predicted as `NIL` (*empty semantics*) 1,022 times, `HAS` (*possessive pronoun*) is predicted as `NIL` 331 times, and `NIL` is predicted as `DEF` (*definite*) 557 times.

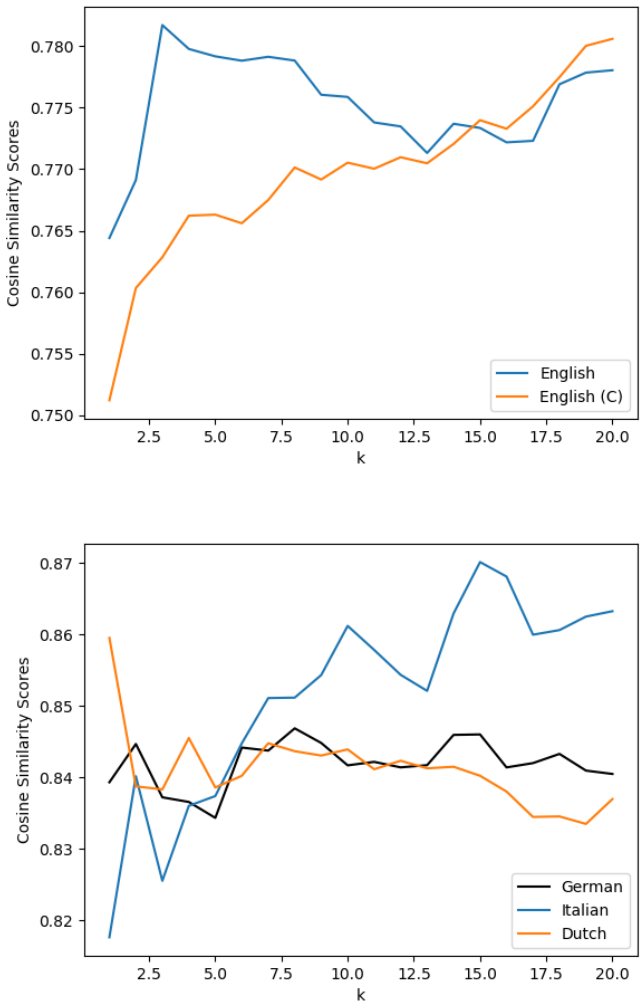

Figure 2: Cosine similarity score vs. k for POS-aware Prediction (top: English, bottom: cross-lingual)

4.5. Generating a Large Dataset

To create a large semantic tagging resource for English words, we apply POS-aware Prediction with $k = 3$ for all words in the vocabulary of the Sketch Engine POS-specific word embeddings. As a result, we obtain semantic tag vectors for 6,143,073 word forms from their English vocabulary. Note that a significant portion of these are named entities. However, the Universal Semantic Tag scheme explicitly considers a sizeable number of different categories of named entities, and in Table 4, we saw that the prediction quality for them (`NAM`) is fairly high.

For our cross-lingual tagging, we apply the same approach for all words in our Wiktionary-based translation data. Our resulting resource contains a total of 370 languages and 1,523,736 word forms with valid semantic tag vectors. Among the 370 languages, 121 have at least 1,000 word forms and 61 have 5,000. The languages with the largest coverage are given in Table 6. We previously saw that cross-lingual prediction works fairly well using the POS-aware Prediction approach in conjunction with a POS-specific translation dictionary. While the set of languages that we operate on here is much more typologically diverse than the 4 Indo-European languages considered in Table 3, we conjecture that the quality depends primarily on the accuracy of the translation resource (de Melo and Weikum, 2009).

5. Conclusion

Universal Semantic Tags are a promising new way of comprehensively labeling words with regard to salient semantic characteristics, making fine-grained distinctions neglected in other tagging schemes.

Our work shows the feasibility of predicting tag distribution vectors for unseen words. We induce a lexicon of Universal Semantic Tag vectors for a large set of word forms both in English and across many other languages. Abdou et al. (2018) demonstrated the usefulness of such semantic tags in several downstream tasks via multi-task learning, including on the Stanford NLI corpus, SICK, POS tagging, and dependency tagging. Hence, we envision our data being useful in a wide range of tasks that benefit from semantic information about words. Our lexical data is available for download from `http://semantictags.nlproc.org/`.

6. References

Abdou, M., Kulmizev, A., Ravishankar, V., Abzianidze, L., and Bos, J. (2018). What can we learn from semantic tagging? In *Proceedings of EMNLP 2018*, pages 4881–4889, Brussels, Belgium. Association for Computational Linguistics.

Abzianidze, L. and Bos, J. (2017). Towards universal semantic tagging. In *Proceedings of IWCS 2017 – 12th International Conference on Computational Semantics*.

Abzianidze, L., Bjerva, J., Evang, K., Haagsma, H., van Noord, R., Ludmann, P., Nguyen, D.-D., and Bos, J. (2017). The Parallel Meaning Bank: Towards a multilingual corpus of translations annotated with compositional meaning representations. In *Proceedings of EACL 2017*, pages 242–247, Valencia, Spain. Association for Computational Linguistics.

Bjerva, J., Plank, B., and Bos, J. (2016). Semantic tagging with deep residual networks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3531–3541.

Brants, T. (2000). TnT – a statistical part-of-speech tagger. In *Sixth Applied Natural Language Processing Conference*, pages 224–231, Seattle, Washington, USA. Association for Computational Linguistics.

Cucerzan, S. and Yarowsky, D. (2000). Language independent, minimally supervised induction of lexical probabilities. In *Proceedings of ACL 2000*, pages 270–277, Hong Kong. Association for Computational Linguistics.

de Melo, G. and Weikum, G. (2009). Towards a universal wordnet by learning from combined evidence. In David Wai-Lok Cheung, et al., editors, *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM 2009)*, pages 513–522, New York, NY, USA. ACM.

de Melo, G. (2014). Etymological Wordnet: Tracing the history of words. In Nicoletta Calzolari, et al., editors, *Proceedings of the 9th Language Resources and Evaluation Conference (LREC 2014)*, pages 1148–1154, Paris, France. European Language Resources Association (ELRA).

de Melo, G. (2015). Wiktionary-based word embeddings. In *Proceedings of MT Summit XV*, pages 346–359. AMTA.

de Melo, G. (2017a). Inducing conceptual embedding spaces from Wikipedia. In *Proceedings of WWW 2017*, pages 43–50. ACM.

de Melo, G. (2017b). Multilingual vector representations of words, sentences, and documents. In *Proceedings of IJCNLP 2017*.

Dong, X. and de Melo, G. (2018a). Cross-lingual propagation for deep sentiment analysis. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI 2018)*, pages 5771–5778. AAAI Press.

Dong, X. and de Melo, G. (2018b). A helping hand: Transfer learning for deep sentiment analysis. In *Proceedings of ACL 2018*, pages 2524–2534.

Faruqui, M. and Dyer, C. (2015). Non-distributional word vector representations. In *Proceedings of ACL-ICJNLP 2015*, pages 464–469. Association for Computational Linguistics.

Fu, Z., Xian, Y., Geng, S., Ge, Y., Wang, Y., Dong, X., Wang, G., and de Melo, G. (2020). Absent: Cross-lingual sentence representation mapping with bidirectional gans. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI 2020)*. AAAI Press.

Mikolov, T., Le, Q. V., and Sutskever, I. (2013). Exploiting similarities among languages for machine translation. *arXiv*, 1309.4168.

Pennington, J., Socher, R., and Manning, C. D. (2014). GloVe: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Shoeb, A. A. M., Raji, S., and de Melo, G. (2019). EmoTag – Towards an emotion-based analysis of emojis. In *Proceedings of RANLP 2019*, pages 1094–1103.