Distributional Representation Clusters Complement Part-of-Speech Tags

Anonymous EMNLP submission

Abstract

Many works have successfully co-opted word clusters derived from distributional information, such as Brown clusters, as features in language processing tasks. We note that not only do such clusters make poor proxies for part-of-speech tags; these clusters are in fact quite different from part-of-speech tags. This paper investigates the gap between Brown clusters, clusterings in word embedding space, and part-of-speech tags, across a range of languages. We find that, while word types clustered together may seem at a glance to be cohesive, distributionally derived clusters in fact strongly complement part-of-speech tags across many languages, suggesting a surprising amount of difference between the information contained in these representations.

1 Introduction

Despite common wisdom, there is an absence of evidence that distributionally-generated word clusters correspond to part-of-speech tags. Indeed, Brown clusters (for example) often outperform other techniques in unsupervised part-of-speech tagging, providing strong prototypes for tag classes (Christodoulopoulos et al., 2010). The research presented in this paper is an empirical approach to demonstrating that such distributional clusters have little to do with, and in fact complement, parts of speech.

Distributionally derived clusters do play an important role in contemporary NLP. Early work focused on class-based models for machine translations (Brown et al., 1992). Shortly after, such cluster were found to be helpful in parsing (Magerman, 1995; Koo et al., 2008) and named entity recognition (Miller et al., 2004; Turian et al., 2010). More recently, Mayhew et al. (2017) found that Brown cluster features remain an important signal for cross-lingual named entity recognition (NER), and Ling et al. (2016) find them even stronger than continuous-bag-of-words vectors for standard NER. Indeed, Brown clusters outperform word vectors and also structural correspondence learning for historical English (Yang and Eisenstein, 2016), and tend to learn more helpful representations using the same amount of data than embeddings (Qu et al., 2015). Finally, while Blunsom and Cohn (2011)'s Pitman-Yor uses distributionally-derived clusters for unsupervised PoS induction and they help, the clusters do not appear to behave as if they are strong predictors of word class.

Our hypothesis is, then, that distributional clusters complement part-of-speech. We test the hypothesis by performing various word clusterings in many languages and comparing them to automatically induced word embeddings. We use a variety of metrics to compare clusters of word embeddings to clusters grouped by part of speech. Low similarity scores between Brown clusters or clusters of embeddings, compared to part-of-speech labels, indicate a dissimilarity. Finding such a dissimilarity tells us that distributional information used in these ways is distinct from part-of-speech, offering a useful complementarity.

2 Method

We select various, diverse languages from the UD Treebank (Nivre et al., 2017), and use these corpora for our experiments. These are sampled in order to correct for cross-linguistic variation. Distributional representations are then derived from these samples. We convert these representations to clusters, the same number as there are part-of-speech tags used – seventeen for UD – and then compare how closely each distributionally-derived clustering compares with part of speech tags.

2.1 Data Sampling

Representation inductions have sometimes been performed on larger corpora in the past. We reduce this in order to examine a broad range of languages while also using a comparable amount of data for each language. The languages we select are Danish, English, Finnish, French, Hebrew, Portuguese, Russian, Urdu, and Chinese, to represent a somewhat diverse set of language families. Choosing corpus size fairly across languages is non-trivial. Each token represents a different amount of information; agglutinative languages may express in one word what others use a whole sentence to achieve. For example, from Finnish:

> *Juoksentelisinkohan* (1 token)
> *I wonder if I should run around aimlessly?* 9 tokens

One principled way to correct for this is to use multi-text, i.e. k-way translations of the same content (Cotterell et al., 2018). This allows calculation of BPEC (bits per English character) which standardises across orthographic or phonological variation. However if we are to examine a broad range of languages, we are not immediately afforded this luxury of translated resources, as one might be if studying European languages alone (through EuroParl). As secondary measure, then, we normalise corpus size by bits per character (BPC), defined as $\frac{1}{|c|+1}\sum_{i=1}^{|c|+1}\log p(c_i|\mathbf{c}_{<i})$, where single characters c are characters in an observed corpus (Cotterell et al., 2018).

Values for BPC and the resulting corpus sizes are shown in Table 1. Figures for Europarl languages are taken from (Cotterell et al., 2018). These are then linked via entropy estimates, , to non-Europarl languages, using data from Kolmogorov (1965), Khan et al. (1984), Chang and Lin (1994) and Levitin and Reingold (1994). As Europarl is a single-genre dataset and so not extraordinarily diverse, we calibrate these BPC figures using a general estimate for the entropy of English of 1.46 (Teahan and Cleary, 1996). This allows selection of datasets having very close to the exact same number of non-punctuation non-space characters, giving cross-language data normalised for information content.

2.2 Brown Clusters

Brown clustering (Brown et al., 1992) is a hierarchical hard clustering algorithm that uses decrease in global aggregate mutual information (AMI) as the loss metric. The two clusters that, when merged, cause the least loss in global mutual information, are merged at each step. As the search space here is large and needs to be partly recomputed each merge, it is typical to constrain it to a set breadth, a; often 1000. Our aim is 17 clusters (corresponding to the number of UD PoS tags). Considering only the top 17 most-frequent terms at each merge gives a narrow window and leads to a high AMI loss and mostly sub-optimal merges. Therefore, using the generalised formulation of the Brown clustering algorithm, we consider 2500 clusters at each merge, and thereafter use roll-up feature extraction (Derczynski and Chester, 2016) to get the final 17; a comparison of this versus setting $a = 17$ is also provided.

Language	EuroParl BPC	General BPC	Tokens
Danish	1.11	1.47*	80K
English	1.10	1.46	83K
Finnish	1.16	1.54*	65K
French	0.95	1.26*	95K
Hebrew	1.11*	1.47	70K
Portuguese	1.01	1.34*	91K
Russian	0.87*	1.15	67K
Urdu	1.38*	1.84	52K
Chinese	2.92*	3.88	31K

Table 1: Sizes of corpora normalised by bits per character. * = scaled figure.

2.3 Embedding Clusters

Word embeddings are representations of words in vector space. We run GloVe (Pennington et al., 2014) over each of our scaled corpora to induce these vector representations. Then, we cluster the vector representations, creating the same number of clusters (17) as there are part-of-speech tags.

We need to arrive at a number of contiguous clusters from this representation that completely cover all word types. As some *clustering* algorithms ignore outlying points, leaving them without a label, instead a *partitioning* algorithm is required, which will completely cover the data. We use k-means (Lloyd, 1982) for this.

Typical partitioning algorithms, including k-means, tend to underperform in very high-dimensional space, because they rely on Euclidean (L2) distance. L2 becomes meaningless in high dimensions; due to small variations compounded over multiple dimensions, every point tends toward equidistant (Beyer et al., 1999). By extension, clustering methods that rely on L2 also become meaningless in higher dimensions. In addition, embeddings in lower-dimensional space should converge to a stable state faster, and we are constrained to modest corpus sizes for the sake of covering a broad range of languages (Section 2.1). The only GloVe hyperparameter to adjust is the number of dimensions output vectors should have, and so when building our embeddings, we select 10 dimensions (a relatively low number in the context of word vectors).

2.4 Metric Selection

We evaluate by comparing similarity of clusterings. Many options are available, and so we set specific desiderata. -measure (Rosenberg and Hirschberg, 2007), the harmonic mean of cluster homogeneity and cluster completeness, works well; however, it is known to have a bias toward giving higher scores with higher numbers of clusters (Vinh et al., 2010). Rand Index (Rand, 1971) is another option, though this does not correct for the low baseline level of overlap in random clustering (i.e. is does not exhibit the *constant baseline property*). This can be corrected for by using the adjusted Rand Index, ARI (Steinley, 2004). However, the distance in ARI is not a proper metric, leaving it poor for comparisons in the space of clusterings. For this work therefore we use adjusted mutual information as the cluster similarity metric (Vinh et al., 2010).

Where $\mathbf{U}$ and $\mathbf{V}$ are two clusterings; H is the entropy, and I is the information:

$$AdjMI_{max}(\mathbf{U},\mathbf{V}) = \frac{I(\mathbf{U},\mathbf{V}) - E\{I(\mathbf{U},\mathbf{V})\}}{\max\{H(\mathbf{U}),H(\mathbf{V})\} - E\{I(\mathbf{U},\mathbf{V})\}}$$

Following Hubert and Arabie (1985) in applying chance adjustment, this information-theoretic metric gives the random baseline for free, while simultaneously addressing problems with other popular clustering similarity metrics. Its range is [0..1], where 0 indicates a no-better-than-random clustering similarity and 1 is a perfect overlap.

2.5 Polysemy

We should handle polysemy; in many languages, many word types have more than one possible PoS tag. To handle this, we build clusterings over data where the surface form and instance PoS tag are concatenated. E.g. for the Danish word *ham* occurring as a pronoun, we use the token *ham_PRON*; this produces a unit at slightly coarser than lexeme level, distinguishing a subset of senses for a given word type. Note that this diverges from many of the unsupervised PoS induction methods proposed, including Van Gael et al. (2009).

Language	Brown *a=17*	Brown *a=2.5K*	GloVe
Danish	0.090	0.089	0.117
English	0.093	0.135	0.124
Finnish	0.030	0.022	0.043
French	0.142	0.092	0.171
Hebrew	0.166	0.204	0.085
Russian	0.054	0.067	0.083
Urdu	0.122	0.147	0.111
Chinese	0.081	0.084	0.098

Table 2: Part-of-speech complementarity: cluster similarity between PoS tags and distributionally-derived clusters, measured by adjusted mutual information.

3 Results and Analysis

Table 2 presents cluster similarity with part-of-speech tags. It presents low figures for cluster similarity, indicating that distributionally-derived clusters, both Brown and k-means over word vectors, complement part-of-speech tags. The values for cluster overlap are very low, coming close to random (0). This indicates that distributionally derived word clusters group words very differently from how part of speech tag does, supporting our initial hypothesis.

3.1 Typological Comparison

The results vary between languages. Note that Brown clusters are closer to parts of speech for some languages (Hebrew) than others (Russian). While, given Russian's rich morphology and its frequent expression of grammar through inflection instead of auxiliaries, one might suspect that this language would not lend itself to ready analysis through bigram-based distributional representations, it is a little surprising that Hebrew appears to do so considerably more. The two languages share some grammatical features, such as no requirement for articles before nouns and flexible word order. We see also that clustered GloVe vectors connect well with French parts of speech; French has a somewhat strict word order, which is also used to mark case. Note that GloVe takes bigrams from further than immediate neighbours (i.e. skip-grams), allowing some capture of syntax depending on the breadth of the window – unlike Brown clustering.

Language	Adjusted MI
Danish	0.090
English	0.083
Finnish	0.037
French	0.093
Hebrew	0.053
Russian	0.100
Urdu	0.097
Chinese	0.081

Table 3: Comparison of Brown ($a = 2500$) with GloVe/k-means clusters.

#clust.	node	item	#clust.	node	item
17	0.118	0.001	17	0.474	0.647
100	0.036	0.250	100	0.496	0.390
300	0.064	0.233	300	0.498	0.333
800	0.103	0.281	800	0.547	0.425
1600	0.143	0.325	1600	0.562	0.456

Table 4: Node- and item-level homogeneity in English (left) and Hebrew (right). Hebrew has one of the least dissimilar Brown clusters from part-of-speech ground truth. $h = 0.85$, $a = 2500$.

3.2 Brown Clusters vs. Embedding Clusters

We compare Brown clusters with those built from the GloVe embeddings we derived. Results are in Table 3. Here we can see that Brown clusters and embedding-derived clusters are strongly different, coming close to random in their similarity. The difference is particularly strong for Hebrew, where there was also a relatively strong difference in cluster similarity to part-of-speech tag. This offers empirical evidence to support the folk knowledge from extrinsic evaluation that Brown clusters are an effect complement to embeddings; they offer complementary information.

3.3 Tree Structure Analysis

The hierarchicality of Brown clustering provides an extra level of detail, derived entirely from distributional information. Based on candid examinations of Brown clusters, we might hypothesize that words of the same PoS accumulate in individual clusters or subtrees, as have others in the past (Yang and Eisenstein, 2016). Indeed, distributionally derived clusters can appear to contain similar words at a glance; one may readily find clusters exclusively representing phenomena such as months of the year, days of the week, synonyms for "good", or spelling variations of the word "tomorrow" (Ritter et al., 2011).

To analyse this observation, we iteratively expand the tree, unrolling it in reverse merge order, and measure the *homogeneity* of part-of-speech of each node (Rosenberg and Hirschberg, 2007). If it is that words with the same part-of-speech are placed in the same distributional cluster, there will be high-homogeneity groups. For example, while high-level nodes are likely to comprise a broad range of words and classes, it is possible that nodes deeper down the tree are by dominated by one single part-of-speech.

To measure this, we set a threshold h for a minimum homogeneity of part-of-speech tag that a node may have. The extent of homogeneity is then calculated two ways. Firstly, node homogeneity: the proportion of all nodes that are homogeneous, i.e. the dominant part of speech accounts for more that h of the group. Secondly, item homogeneity: how many items are accounted for by homogeneous groups to account for the volume of words in each group. Note that we refer to item instead of word type as words are split by part of speech (Section 2.5). For example, a group consisting of Danish { *ham_PRON*, *de_PRON* } is homogeneous regarding part of speech, but only accounts for two items, and so contributes to the first metric as much as a very large homogeneous node, but less under the second metric.

We can see that, interestingly, as we "unfurl" the clustering down its hierarchy, there is a paucity of clusters that are homogeneous in terms of part-of-speech tag for English. The scores here are consistently low. In contrast, Hebrew Brown clusters relate more strongly to part-of-speech than English ones. With 17 clusters, almost half the nodes in the Brown tree have over 85% of their members being the same part of speech.

4 Conclusions

Distributionally-derived clusters are distinct from part-of-speech tags. Indeed, distributional similarity does not directly predict part-of-speech tag; rather, the information is largely complementary, with the extent varying across languages. The representations from Brown clustering and from partitioning words in embedding space are complementary; so, both should be experimented with as features and sources of information. In closing – these clusterings relate differently across different language types, have internal cohesiveness, and have been found by many others to be useful in language processing. This motivates an interesting avenue of investigation: what do the clusters actually mean?

References

Kevin Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft. 1999. When is "nearest neighbor" meaningful? In *International conference on database theory*, pages 217–235. Springer.

Phil Blunsom and Trevor Cohn. 2011. A hierarchical Pitman-Yor process HMM for unsupervised part of speech induction. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 865–874. Association for Computational Linguistics.

Peter F Brown, Peter V Desouza, Robert L Mercer, Vincent J Della Pietra, and Jenifer C Lai. 1992. Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–479.

Jyun-sheng Chang and Yuh-Juh Lin. 1994. An estimation of the entropy of Chinese - A new approach to constructing class-based n-gram models. In *Proceedings of Rocling VII Computational Linguistics Conference VII*, pages 149–169. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Christos Christodoulopoulos, Sharon Goldwater, and Mark Steedman. 2010. Two decades of unsupervised POS induction: How far have we come? In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 575–584. Association for Computational Linguistics.

Ryan Cotterell, Sebastian J. Mielke, Jason Eisner, and Brian Roark. 2018. Are all languages equally hard to language-model? In *Proceedings of NAACL-HLT*.

Leon Derczynski and Sean Chester. 2016. Generalised Brown Clustering and Roll-Up Feature Generation. In *AAAI*, pages 1533–1539.

Lawrence Hubert and Phipps Arabie. 1985. Comparing partitions. *Journal of classification*, 2(1):193–218.

MZ Khan, M Ilyas, and I Saud. 1984. Entropy and internal information of printed Urdu. *Journal of Applied Statistics*, 11(2):142–154.

Andrei N Kolmogorov. 1965. Three approaches to the quantitative definition of information. *Problems of information transmission*, 1(1):1–7.

Terry Koo, Xavier Carreras, and Michael Collins. 2008. Simple semi-supervised dependency parsing. *Proceedings of ACL-08: HLT*, pages 595–603.

Lev B Levitin and Zeev Reingold. 1994. Entropy of natural languages: Theory and experiment. *Chaos, Solitons & Fractals*, 4(5):709–743.

Shaoshi Ling, Yangqiu Song, and Dan Roth. 2016. Word embeddings with limited memory. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 387–392.

Stuart Lloyd. 1982. Least squares quantization in PCM. *IEEE transactions on information theory*, 28(2):129–137.

David M Magerman. 1995. Statistical decision-tree models for parsing. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, pages 276–283. Association for Computational Linguistics.

Stephen Mayhew, Chen-Tse Tsai, and Dan Roth. 2017. Cheap translation for cross-lingual named entity recognition. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2536–2545.

Scott Miller, Jethran Guinness, and Alex Zamanian. 2004. Name tagging with word clusters and discriminative training. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*.

Joakim Nivre, Željko Agić, Lars Ahrenberg, et al. 2017. Universal Dependencies 2.1. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Lizhen Qu, Gabriela Ferraro, Liyuan Zhou, Weiwei Hou, Nathan Schneider, and Timothy Baldwin. 2015. Big Data Small Data, In Domain Out-of Domain, Known Word Unknown Word: The Impact of Word Representations on Sequence Labelling Tasks. *CoNLL 2015*, page 83.

William M Rand. 1971. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850.

Alan Ritter, Sam Clark, Oren Etzioni, et al. 2011. Named entity recognition in tweets: an experimental study. In *Proceedings of the conference on empirical methods in natural language processing*, pages 1524–1534. Association for Computational Linguistics.

Andrew Rosenberg and Julia Hirschberg. 2007. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*.

Douglas Steinley. 2004. Properties of the Hubert-Arabie Adjusted Rand Index. *Psychological methods*, 9(3):386.

William J Teahan and John G Cleary. 1996. The entropy of English using PPM-based models. In *Data Compression Conference, 1996. DCC'96. Proceedings*, pages 53–62. IEEE.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 384–394. Association for Computational Linguistics.

Jurgen Van Gael, Andreas Vlachos, and Zoubin Ghahramani. 2009. The infinite HMM for unsupervised PoS tagging. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*, pages 678–687. Association for Computational Linguistics.

Nguyen Xuan Vinh, Julien Epps, and James Bailey. 2010. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11(Oct):2837–2854.

Yi Yang and Jacob Eisenstein. 2016. Part-of-Speech Tagging for Historical English. In *Proceedings of NAACL-HLT*, pages 1318–1328.