Unsupervised Language Learning - Practical 1

Nuno Mota - 11413344, Tom Pelsmaeker - 10177590

April 24, 2018

In this report and implementation[a], dependency-based (deps), bow-2 and bow-5 word embeddings are compared on their relative performance on word similarity tasks, a word analogy task and clustering. The results and implications will be discussed below.

Word Similarity Tasks

	Dependency	BOW-2	BOW-5
SimLex	0.462/0.446	0.428/0.414	0.376/0.367
MEN	0.597/0.618	0.678/0.700	0.708/0.723

Table 1: Pearson/Spearman correlation of various word embedding matrices on similarity task.

Table 1 shows that deps performs best on SimLex, whereas bow-5 performs best on MEN. This is expectable as MEN measures *relatedness* (e.g. vehicle-car) and SimLex *similarity* (e.g. automobile-car). Words in the same context, as captured by bow, are often related, but not necessarily similar. Words that have the same function, as captured by deps, are often similar. MEN is easier in general as similarity can be seen as a special case of relatedness. The word pairs with highest cosine similarity of the three embedding types display this similarity vs relatedness dichotomy (Appendix Table 3).

Word Analogy Task

The cosine similarity on this task was found as the inner product of the normalised embeddings. Vectors were then ordered by highest inner product value. The source vector ”b” had to be removed from this ordering to obtain good accuracy and MRR.

	Acc $\left(\cup \vec{b}\right)$%	MRR $\left(\cup \vec{b}\right)$
Dependency	27.07 (3.53)	0.38 (0.24)
BOW-2	42.09 (9.32)	0.57 (0.37)
BOW-5	54.67 (10.45)	0.66 (0.41)

Table 2: Metrics on the word analogy task, for the lowercased dataset and when the b (from $a : a^* :: b : b^*$) word is disregarded (or not) when computing the metrics.

As the analogy task tests accuracy of embeddings in encoding 19 different types of relations it is perhaps unsurprising that the bow embeddings outperform deps. Such analogies can be seen as comparing certain types of relatedness between words, which, we already saw on the previous task, is captured in a broader manner by bow embeddings than by deps embeddings.

Additionally, observing Table 4 (Appendix), we can qualitatively verify, on a few random examples, that the dependency dataset seems indeed to better capture words with similar function, while the BOW datasets capture more related words. Focusing, for example, on the ”moscow –> russia” relation, we can see that the dependency dataset focuses more on countries, in general, and the bow datasets focus more on Russian related words, such as Leningrad, Irkutsk and former USSR countries, which seems to follow the expected behavior.

Clustering

A set of 2000 noun embeddings were visualized in 2-D with t-SNE (Appendix Figure 1), showing a few dense clusters but mostly tiny cluster within a single large cluster. Perhaps, general 'noun-ness' was captured in the large cluster, with special relations between nouns being captured in the small dense clusters.

Clustering was then performed using density-based DBSCAN clustering with cosine similarity as metric. This method is preferred over K-means as it does not force all the data to be clustered, only clustering strongly similar words. Further, it does not require the specification of the number of clusters, which is unknown in our case.

In general, the clusters of the three different embeddings are comparable. They all contain a cluster of names, jobs, countries, and many bi-word clusters of closely related words. However, deps yields large clusters of a certain concept, e.g. jobs, sentiments, money, that are less apparent in bow-5 and bow-2. This may be because such concepts are often found in the same dependency structure, but not necessarily in the same local context. For instance, jobs will often occur in similar structure, e.g. 'the poet/politician worked on ...'. Yet the context may differ, a politician will work on something else than a poet. Some clusters can be found in Appendix Table 5.

Conclusion

The three types of word embeddings show different properties; deps generally embeds functionally similar words close, whereas bow-5 embeds topically similar words close. Bow-2 falls between the two other types, encoding less topicality than bow-5 due to the smaller window, but also less functionality as it cannot access the informative dependency structure.

[a] `https://github.com/0Gemini0/ULL/tree/master/Practical1`

Appendix

	Deps	Bow-2	Bow-5
SimLex	actress-actor	south-north	sheep-cattle
	south-north	sheep-cattle	actress-actor
	archbishop-bishop	actress-actor	south-north
	delightful-cheerful	movie-film	archbishop-bishop
	movie-film	archbishop-bishop	winter-summer
MEN	amphibians-reptiles	amphibians-reptiles	boys-girls
	mammals-reptiles	boys-girls	cattle-sheep
	boys-girls	cattle-sheep	carrots-potatoes
	amphibians-mammals	mammals-reptiles	mammals-reptiles
	carrots-potatoes	bicycle-bike	beef-meat

Table 3. Top-5 most similar word pairs on SimLex and MEN rated by cosine similarity of the three word embedding types. On the SimLex dataset it can be seen that the deps embeddings encode similarity better than bow-2 or bow-5, which both rank the related but not similar sheep-cattle word pair in the top 5. The deps ranking does however contain the south-north pair in the top 5, which are related but not similar. This illustrates that the functional dependency captured by the deps embeddings can also coincide between related words.

Example	Dependency	BOW-2	BOW-5
loud –> louder	louder, safer, quieter, funnier	squealing, louder, shrill, grunting	louder, shrill, high-pitched, squealing
moscow –> russia	russia, ethiopia, uzbekistan, tunisia	russia, irkutsk, tajikistan, ukraine	russia, leningrad, ukraine, belarus
bird –> birds	birds, rats, frogs, butterflies	birds, songbirds, seabirds, dragonflies	birds, rats, rabbits, waterfowl
code –> coding	codes, coding, identifier, six-digit	codes, 3166-1, 639-1, two-letter	codes, two-letter, 639-1, 10-digit

Table 4. Comparisons on the word analogy task. Choice of (a : a :: b : b*) was random. The table presents the top 4 choices, for each dataset, excluding the "b" word. We can indeed observe that with Deps we capture more the words with similar function, while with the Bow datasets we capture more related words.*

Deps	Bow-2	Bow-5
'ability', 'ambition', 'anger', 'anxiety', 'capability', 'commitment', 'confusion', 'contribution', 'curiosity', 'desire', 'disappointment', 'embarrassment', 'emotion', 'enthusiasm', 'excitement', ...	'commitment', 'desire', 'refusal', 'willingness'	'ability', 'ambition', 'desire', 'refusal', 'willingness'
'accountant', 'adviser', 'analyst', 'applicant', 'architect', 'artist', 'auditor', 'author', 'buyer', 'captain', 'chairman', 'clerk', 'client', 'coach', 'commander', ...	'analyst', 'author', 'commentator', 'critic', 'historian', 'journalist', 'photographer', 'poet', 'reporter', 'writer'	'adviser', 'analyst', 'artist', 'author', 'commentator', 'consultant', 'critic', 'historian', 'journalist', 'lawyer', 'photographer', 'poet', 'politician', 'reporter', 'writer'
'alice', 'anna', 'anne', 'athelstan', 'blanche', 'caroline', 'charles', 'charlie', 'diana', 'edward', 'elizabeth', 'emily', 'francis', 'george', 'harry', ...	'charles', 'edward', 'francis', 'george', 'henry', 'james', 'john', 'richard', 'robert', 'thomas', 'william'	'adam', 'baker', 'benjamin', 'charles', 'charlie', 'clarke', 'david', 'edward', 'francis', 'george', 'graham', 'harry', 'henry', 'howard', 'jack', ...
'advice', 'guidance', 'proposal', 'recommendation', 'request', 'suggestion'	'emily', 'helen', 'jane', 'kate', 'laura', 'lucy', 'maggie', 'rachel', 'ruth', 'sarah', 'susan'	alice', 'anna', 'anne', 'blanche', 'caroline', 'elizabeth', 'emily', 'helen', 'jane', 'kate', 'laura', 'lucy', 'maggie', 'marie', 'mary', 'rachel', 'ruth', 'sarah', 'susan'
'allowance', 'cheque', 'compensation', 'dividend', 'earnings', 'expenditure', 'income', 'lease', 'loan', 'mortgage', 'payment', 'pension', 'receipt', 'revenue', 'salary', 'subsidy', 'wages'	'east', 'north', 'south', 'west'	'leeds', 'liverpool', 'london'
'assumption', 'concept', 'expectation', 'idea', 'implication', 'notion', 'premise', 'principle'	'expertise', 'knowledge', 'understanding'	expertise', 'knowledge', 'understanding'
'baby', 'brother', 'child', 'colleague', 'cousin', 'daughter', 'father', 'friend', 'husband', 'lover', 'mother', 'sister', 'wife'	'brother', 'daughter', 'father', 'husband', 'mother', 'wife'	'brother', 'colleague', 'cousin', 'daughter', 'father', 'friend', 'husband', 'lover', 'mother', 'sister', 'wife'
'beer', 'bread', 'cake', 'champagne', 'coffee', 'cream', 'food', 'fruit', 'grain', 'juice', 'meat', 'milk', 'potato', 'sugar', 'wine'	'britain', 'france', 'germany', 'italy', 'spain'	'bread', 'breakfast', 'cake', 'cream', 'dinner', 'fruit', 'juice', 'lunch', 'meal', 'meat', 'milk', 'potato', 'sugar'
'accident', 'crash', 'incident'	'enquiry', 'inquiry', 'investigation'	'asset', 'investment', 'investor'

Table 5. Selection of clusters found with DBSCAN. When possible, similar clusters are on the same line across embedding types. Full clusters can be reproduced with the code on GitHub. Note that deps clusters all names together, whereas bow can discriminate between women's and men's names. Female names have different context but similar dependency compared to male names.

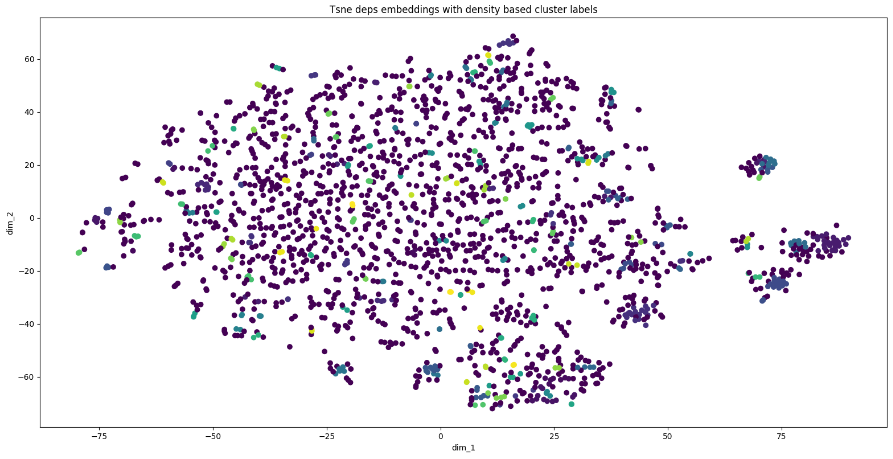

Figure 1. T-SNE reduced dependency word embeddings of 2000 nouns. Clusters found with DBSCAN are superimposed on the image to show general agreement between the visualization and found clusters. With K-Means clustering no such agreement was apparent. Specifically note the large clusters in the remote parts of the visualization. Visualization of the other two embedding types was similar, and is thus left out. Note that the clusters were computed with unreduced embeddings.