Evaluating embeddings on dictionary-based similarity

Judit Ács
Department of Automation
Budapest University of Technology
Magyar Tudósok krt 2
1111 Budapest, Hungary
judit@aut.bme.hu

András Kornai
Institute for Computer Science
Hungarian Academy of Sciences
Kende u. 13-17
1111 Budapest, Hungary
andras@kornai.com

Abstract

We propose a method for evaluating embeddings against dictionaries with tens or hundreds of thousands of entries, covering the entire gamut of the vocabulary.

1 Introduction

Continuous vector representations (embeddings) are, to a remarkable extent, supplementing and potentially taking over the role of detail dictionaries in a broad variety of tasks ranging from POS tagging (Collobert et al., 2011) and parsing (Socher et al., 2013) to MT (Zou et al., 2013), and beyond (Karpathy, Joulin, and Li, 2014). Yet an evaluation method that directly compares embeddings on their ability to handle word similarity at the entire breadth of a dictionary has been lacking, which is all the more regrettable in light of the fact that embeddings are normally generated from gigaword or larger corpora, while the state of the art test sets surveyed in Chiu, Korhonen, and Pyysalo (2016) range between a low of 30 (MC-30) and a high of 3,000 word pairs (MEN).

We propose to develop a dictionary-based standard in two steps. First, given a dictionary such as the freely available Collins-COBUILD (Sinclair, 1987), which has over 77,400 headwords, or Wiktionary (162,400 headwords), we compute a frequency list F that lists the probabilities of the headwords (this is standard, and discussed only briefly), and a dense similarity matrix M or an embedding ψ, this is discussed in Section 2. Next, in Section 3 we consider an arbitrary embedding ϕ, and we systematically compare both its frequency and its similarity predictions to the gold standard embodied in F and ψ, building on the insights of Arora et al. (2015). Pilot studies conducted along these lines are discussed in Section 4.

Before turning to the details, in the rest of this Introduction we attempt to evaluate the proposed evaluation itself, primarily in terms of the criteria listed in the call. As we shall see, our method is *highly replicable for other researchers* for English, and to the extent monolingual dictionaries are available, for other other languages as well. Low resource languages will typically lack a monolingual dictionary, but this is less of a perceptible problem in that they also lack larger corpora so building robust embeddings is already out of the question for these. *The costs are minimal*, since we are just running software on preexisting dictionaries. Initially, dictionaries are hard to assemble, require a great deal of manual labor, and are often copyrighted, but here our point is to leverage the manual (often crowdsourced) work that they already embody.

The proposed algorithm, as we present it here, is aimed primarily at *word-level* evaluation, but there are standard methods for extending these from word to sentence similarity (Han et al., 2013). Perhaps the most attractive *downstream application* we see is MT, in particular word sense disambiguation during translation. As for *linguistic/semantic/psychological properties*, dictionaries, both mono- and bilingual, are crucial resources not only for humans (language learners, translators, etc.) but also for a variety of NLP applications, including MT, cross-lingual information retrieval, cross-lingual QA, computer-assisted language learning, and many more. The mandate of lexicographers is to capture a huge number of linguistic phenomena ranging from gross synonymy to subtle meaning distinctions, and at the semantic level the *inter-annotator agreement is very high*, a point we discuss in greater detail below. Gladkova and Drozd (2016) quote Schütze (2016) that “human linguistic judgments (...) are subject to over 50 potential linguistic, psychologi-

Proceedings of the 1st Workshop on Evaluating Vector Space Representations for NLP, pages 78–82,
Berlin, Germany, August 12, 2016. ©2016 Association for Computational Linguistics

cal, and social confounds", and many of these taint the crowd-sourced dictionaries, but lexicographers are annotators of a highly trained sort, and their work gives us valuable data, as near to laboratory purity as it gets.

2 Constructing the standard

Our main inputs are a frequency list F, ideally generated from a corpus we consider representative of the text of interest (the expected input to the downstream task), and a preexisting dictionary D which is not assumed to be task-specific. For English, we use both the Collins-COBUILD dictionary (CED) and Wiktionary, as these are freely available, but other general-purpose dictionaries would be just as good, and for specific tasks (e.g. medical or legal texts) it may make sense to add in a task-specific dictionary if available. Neither D nor F need contain the other, but we assume that they are stemmed using the same stemmer.

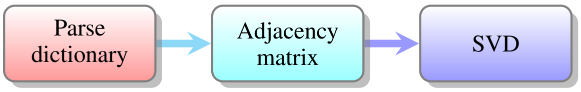

Figure 1: Building the standard

The first step is to parse D into $\langle$word, definition$\rangle$ stanzas. (This step is specific to the dictionary at hand, see e.g. Mark Lieberman's *readme.* for CED). Next, we turn the definitions into dependency graphs. We use the Stanford dependency parser (Chen and Manning, 2014) at this stage, and have not experimented with alternatives. This way, we can assign to each word a graph with dependency labels, see Fig 2 for an example, and Recski (2016) for details. The dependency graphs are not part of the current incarnation of the evaluation method proposed here, but are essential for our future plans of extending the evaluation pipeline (see Section 4).

In the second step we construct two global graphs: the *definitional dependency* graph DD which has a node for each word in the dictionary, and directed edges running from w_i to w_j if w_j appears in the definition of w_i; and the *headword graph* HG which only retains the edge running from the definiendum to the head of the definiens. We take the head to be the 'root' node returned by the Stanford parser, but in many dictionaries the syntactic head of the definition is typographically set aside and can be obtained directly from the raw D.

At first blush it may appear that the results of this process are highly dependent on the choice of D, and perhaps on the choice of the parser as well. Consider the definition of *client* taken from four separate sources: 'someone who gets services or advice from a professional person, company, or organization' (Longman); 'a person who pays a professional person or organization for services' (Webster); 'a person who uses the services or advice of a professional person or organization' (Oxford); 'a person or group that uses the professional advice or services of a lawyer, accountant, advertising agency, architect, etc.' (dictionary.com).

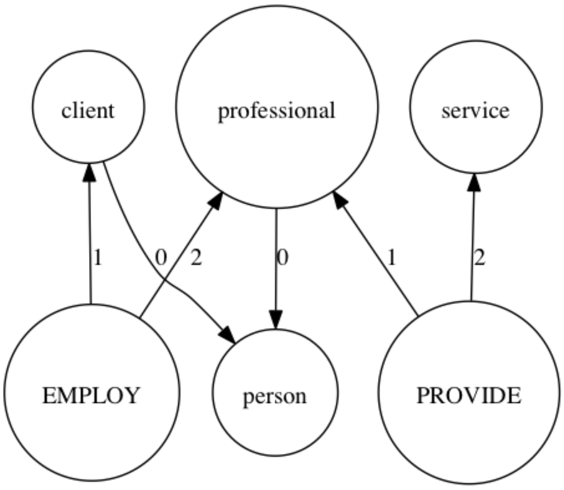

Figure 2: Graph assigned to *client*. Edge labels are 0=isa; 1=nsubj; 2=dobj

The definitions do not literally preserve the headword (hypernym, genus, IS_A): in three cases we have 'person', in one 'somebody'. But semantically, these two headwords are very close synonyms, distinguished more by POS than by content. Similarly, the various definitions do not present the exact same verbal pivot, 'engage/hire/pay for/use the services of', but their semantic relatedness is evident. Finally, there are differences in attachment, e.g. is the service rendered professional, or is the person/organization rendering the service professional? In Section 3 we will present evidence that the proposed method is not overly sensitive to these differences, because the subsequent steps wipe out such subtle distinctions.

In the third step, by performing SVD on the Laplacian of the graphs DD and HG we obtain two embeddings we call the *definitional* and the *head* embedding. For any embedding ψ, a (symmetric, dense) similarity matrix $M_{i,j}$ is given by the cosine similarity of $\psi(w_i)$ and $\psi(w_j)$. Other methods for computing the similarity matrix M are also possible, and the embedding could also be obtained by direct computation, setting the context window of each word to its definition – we defer the discussion of these and similar alternatives to the concluding Section 4.

Now we define the *direct* similarity of two embeddings ϕ and ψ as the average of the (cosine) similarities of the words that occur in both:

$$S(\phi,\psi) = (\sum_w \frac{\phi(w)\psi(w)}{\|\phi(w)\|\|\psi(w)\|})/|D| \qquad (1)$$

It may also make sense to use a frequency-weighted average, since we already have a frequency table F – we return to this matter in Section 3. In and of itself, S is not a very useful measure, in that even random seeding effects are sufficient to destroy similarity between near-identical embeddings, such as could be obtained from two halves of the same corpus. For example, the value of S between 300-dimensional GloVe (Pennington, Socher, and Manning, 2014) embeddings generated from the first and the second halves of the UMBC Webbase (Han et al., 2013) is only 0.0003. But for any two embeddings, it is an easy matter to compute the rotation (orthonormal transform) R and the general linear transform G that would maximize $S(\phi, R(\psi))$ and $S(\phi, G(\psi))$ respectively, and it is these *rotational* resp. *general* similarities S_R and S_G that we will use. For the same embeddings, we obtain $S_R = 0.709, S_G = 0.734$. Note that only S_R is symmetrical between embeddings of the same dimension, for S_G the order of arguments matters.

With this, the essence of our proposal should be clear: we generate ψ from a dictionary, and measure the goodness of an arbitrary embedding ϕ by means of computing S_R or S_G between ϕ and ψ. What remains to be seen is that different dictionary-based embeddings are close to one another, and measure the same thing.

3 Using the standard

In the random walk on context space model of Arora et al. (2015), we expect the log frequency of words to have a simple linear relation to the length of the word vectors:

$$\log(p(w)) = \frac{1}{2d}||\vec{w}||^2 - \log Z \pm o(1) \qquad (2)$$

Kornai and Kracht (2015) compared GloVe to the Google 1T frequency count (Brants and Franz, 2006) and found a correlation of 0.395, with the frequency model failing primarily in distinguishing mid- from low-frequency words. The key insight we take from Arora et al. (2015) is that an embedding is both a model of frequency, whose merit can be tested by direct comparison to F, and a model of cooccurrence, given by $\log p(w, w') = \frac{1}{2d}||\vec{w} + \vec{w'}||^2 - 2\log Z \pm o(1)$.

Needless to say, the ⟨word, definition⟩ stanzas of a dictionary do not constitute a random walk: to the contrary, they amount to statements of semantic, rather than cooccurrence-based, similarity between definiendum and definiens, and this is precisely what makes dictionaries the appropriate yardstick for evaluating embeddings.

State of the art on Simlex-999 was $\rho = 0.64$ (Banjade et al., 2015), obtained by combining many methods and data sources. More recently, Wieting et al. (2015) added paraphrase data to achieve 0.69, and Recski et al. (2016) added dictionary data to get to 0.76. Standard, widely used embeddings used in isolation do not come near this, the best we tested was `GoogleNews-vectors-negative300`, which gets only $\rho = 0.44$; `senna` gets 0.27; and `hpca.2B.200d` gets 0.16, very much in line with the design goals of Simlex-999. The purely dictionary-based embeddings are even worse, the best obtains only $\rho = 0.082$ at 300 dimensions, $\rho = 0.079$ at 30 dimensions.

A heuristic indication of the observation that choice of dictionary will be a secondary factor comes from the fact that dictionary-based embeddings are close to one another. Table 1 shows S_R for three dictionaries, CED, Wikt, and My (not in the public domain). The numbers above the diagonal at 300 dim, below at 30 dim.

	CED	Wikt	My
CED	1.0	.127	.124
Wikt	.169	1.0	.131
My	.202	.168	1.0

Table 1 S_R for dictionary-based embeddings

A more solid indication comes from evaluating embeddings under Simlex-999, under the dictionary-based similarities, and under some other test sets.

emb.tr.dim	SL999	CED	Wikt	MEN	RW	size ∩
GN-vec-neg.300	.442	.078	.044	.770	.508	1825
glove.840B.300	.408	.058	.047	.807	.449	1998
glove.42B.300	.374	.009	.045	.742	.371	2013
glove.6B.300	.360	.065	.127	.734	.389	1782
glove.6B.200	.340	.060	.118	.725	.383	1782
glove.6B.100	.298	.059	.112	.697	.362	1782
senna.300	.270	.052	.098	.568	.385	1138
glove.6B.50	.265	.040	.087	.667	.338	1782
hpca.2B.200	.164	.040	.140	.313	.176	1315

Table 2: Comparing embeddings by Simlex-999, dictionary S_R, MEN, and RareWord

As can be seen, the ρ and S_R numbers largely, though not entirely, move together. This is akin to the astronomers' method of building the 'distance ladder' starting from well-understood measurements (in our case, Simlex-999), and correlating these to the new technique proposed here. While Chiu, Korhonen, and Pyysalo (2016) make a rather compelling case that testsets such as MEN, Mtruk-28, RareWord, and WS353 are not reliable for predicting downstream results, we present here ρ values for the two largest tasks, MEN, with 3,000 word pairs, and RareWord, ideally 2,034, but in practice considerably less, depending on the intersection of the embedding vocabulary with the Rare Word vocabulary (given in the last column of Table 2). We attribute the failure of the lesser test sets, amply demonstrated by Chiu, Korhonen, and Pyysalo (2016), simply to undersampling: a good embedding will have 10^5 or more words, and the idea of assessing the quality on less than 1% simply makes no sense, given the variability of the data. A dictionary-wide evaluation improves this by an order of magnitude or more.

4 Conclusions, further directions

An important aspect of the proposal is the possibility of making better use of F. By optimizing the frequency-weighted rotation we put the emphasis on the function words, which may be very appropriate for some tasks. In other tasks, we may want to simply omit the high frequency words, or give them very low weights. In medical texts we may want to emphasize the words that stand out from the background English frequency counts. To continue with astronomy, the method proposed in this paper is akin to a telescope, which can be pointed at various phenomena.

It is clear from the foregoing that we are offering not a single measurement yardstick but rather a family of these. Lexicographers actually include information that we are only beginning to explore, such as the NSUBJ and DOBJ relations that are also returned in the dependency parse. These can also be built into, or even selectively emphasized, in the similarity matrix M, which would offer a more direct measurement of the potential of individual embeddings in e.g. semantic role labeling tasks. We can also create large-scale systematic evaluations of paraphrase quality, using definitions of the same word coming from different dictionaries – Wieting et al. (2015) already demonstrated the value of paraphrase information on Simlex-999.

We have experimented with headword graphs that retain only the head of a definition, typically the genus. Since the results were very bad, we do not burden the paper with them, but note the following. HGs are very sparse, and SVD doesn't preserve a lot of information from them (the ultimate test of an embedding would be the ability to reconstruct the dictionary relations from the vectors). Even in the best of cases, such as hypernyms derived from WordNet, the relative weight of this information is low (Banjade et al., 2015; Recski et al., 2016). That said, the impact of hypernym/genus on the problem of hubness (Dinu, Lazaridou, and Baroni, 2015) is worth investigating further.

One avenue of research opened up by dictionary-based embeddings is to use not just the definitional dependency graph, but an enriched graph that contains the unification of all definition graphs parsed from the definitions. This will, among other issues, enable the study of *selectional restrictions* (Chomsky, 1965), e.g. that the subject of *elapse* must be a time interval, the object of *drink* must be a liquid, and so on. Such information is routinely encoded in dictionaries. Consider the definition of *wilt* '(of a plant) to become weak and begin to bend towards the ground, or (of a person) to become weaker, tired, or less confident'. To the extent the network derived from the dictionary already contains selectional restriction information, a better fit with the dictionary-based embedding is good news for any downstream task.

References

Arora, Sanjeev et al. (2015). "Random Walks on Context Spaces: Towards an Explanation of the Mysteries of Semantic Word Embeddings". In: *arXiv:1502.03520v1*.

Banjade, Rajendra et al. (2015). "Lemon and Tea Are Not Similar: Measuring Word-to-Word Similarity by Combining Different Methods". In: *Proc. CICLING15*. Ed. by Alexander Gelbukh. Springer, pp. 335–346.

Brants, Thorsten and Alex Franz (2006). *Web 1T 5-gram Version 1*. Philadelphia: Linguistic Data Consortium.

Chen, Danqi and Christopher D Manning (2014). "A Fast and Accurate Dependency Parser using Neural Networks." In: *EMNLP*, pp. 740–750.

Chiu, Billy, Anna Korhonen, and Sampo Pyysalo (2016). "Intrinsic Evaluation of Word Vectors Fails to Predict Extrinsic Performance". In: *Proc. RepEval (this volume)*. Ed. by Omer Levy. ACL.

Chomsky, Noam (1965). *Aspects of the Theory of Syntax*. MIT Press.

Collobert, R. et al. (2011). "Natural Language Processing (Almost) from Scratch". In: *Journal of Machine Learning Research (JMLR)*.

Dinu, Georgiana, Angeliki Lazaridou, and Marco Baroni (2015). "Improving Zero-shot Learning by Mitigating the Hubness Problem". In: *ICLR 2015, Workshop Track*.

Gladkova, Anna and Aleksandr Drozd (2016). "Intrinsic Evaluations of Word Embeddings: What Can We Do Better?" In: *Proc. RepEval (this volume)*. Ed. by Omer Levy. ACL.

Han, Lushan et al. (2013). "UMBC_EBIQUITY-CORE: Semantic textual similarity systems". In: *Proceedings of the 2nd Joint Conference on Lexical and Computational Semantics*, pp. 44–52.

Karpathy, Andrej, Armand Joulin, and Fei Fei F Li (2014). "Deep Fragment Embeddings for Bidirectional Image Sentence Mapping". In: *Advances in Neural Information Processing Systems 27*. Ed. by Z. Ghahramani et al. Curran Associates, Inc., pp. 1889–1897.

Kornai, András and Marcus Kracht (2015). "Lexical Semantics and Model Theory: Together at Last?" In: *Proceedings of the 14th Meeting on the Mathematics of Language (MoL 14)*. Chicago, IL: Association for Computational Linguistics, pp. 51–61.

Pennington, Jeffrey, Richard Socher, and Christopher Manning (2014). "Glove: Global Vectors for Word Representation". In: *Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*.

Recski, Gábor (2016). "Computational methods in semantics". PhD thesis. Eötvös Loránd University, Budapest.

Recski, Gábor et al. (2016). "Measuring semantic similarity of words using concept networks". to appear in RepL4NLP. URL: `http://hlt.bme.hu/en/publ/Recski%5C_2016c`.

Schütze, Carson T. (2016). *The empirical base of linguistics*. 2nd ed. Vol. 2. Classics in Linguistics. Berlin: Language Science Press.

Sinclair, John M. (1987). *Looking up: an account of the COBUILD project in lexical computing*. Collins ELT.

Socher, R. et al. (2013). "Zero-shot learning through cross-modal transfer". In: *International Conference on Learning Representations (ICLR)*.

Wieting, John et al. (2015). "From Paraphrase Database to Compositional Paraphrase Model and Back". In: *TACL* 3, pp. 345–358.

Zou, Will Y et al. (2013). "Bilingual Word Embeddings for Phrase-Based Machine Translation." In: *EMNLP*, pp. 1393–1398.