Towards pertinent evaluation methodologies for word-space models

Magnus Sahlgren

SICS, Swedish Institute of Computer Science
Box 1263, SE-164 29 Kista, Sweden
mange@sics.se

Abstract

This paper discusses evaluation methodologies for a particular kind of meaning models known as word-space models, which use distributional information to assemble geometric representations of meaning similarities. Word-space models have received considerable attention in recent years, and have begun to see employment outside the walls of computational linguistics laboratories. However, the evaluation methodologies of such models remain infantile, and lack efforts at standardization. Very few studies have critically assessed the methodologies used to evaluate word spaces. This paper attempts to fill some of this void. It is the central goal of this paper to answer the question "how can we determine whether a given word space is a *good* word space?"

1. Word-space models

Word-space models (Gallant, 1991b; Schütze, 1993; Lund et al., 1995; Landauer and Dumais, 1997; Sahlgren, 2005) use distributional statistics to acquire representations of word meaning. The underlying hypothesis behind these models is that the distributional profiles of words are symptomatic for their semantic content, and that a geometric representation of these profiles is computationally (and, some would argue) cognitively plausible. Both the distributional hypothesis of word meaning and the geometric representational scheme have proven their mettle in such diverse experimental settings as information retrieval (Deerwester et al., 1990; Gallant, 1991a; Jiang and Littman, 2000), vocabulary tests (Landauer and Dumais, 1997; Karlgren and Sahlgren, 2001), word sense disambiguation (Schütze, 1992), lexical priming tests (Lund and Burgess, 1996; McDonald and Lowe, 1998), text categorization (Sahlgren and Cöster, 2004), and so on. There is certainly no shortage of research results arguing for the viability of the approach.

Much thanks to these experimental testimonies, word-space models are becoming established as part of the basic arsenal of language technology. In addition to purely experimental relevance, there is a growing interest in using word-space models for more practically oriented applications, such as knowledge assessment, information extraction, and spam filtering. Furthermore, word-space models are increasingly used for the automatic construction of language resources. To take but one example, word-space models have been used with greatly encouraging results for acquiring thesauri from raw data (Sahlgren and Karlgren, 2005). Such applications will become ever more common — and useful — in the face of a rapidly expanding and flourishing multilingual, multi-cultural, and multi-ethnic computational linguistics community. Word-space modeling is, to say the least, an active area of research.

2. The need for critical assessment of evaluation methodologies

Despite (or perhaps due to) this optimistic climate, the quality of the evaluation methodologies used in word-space research has not received much attention. This is remarkable, for a number of reasons.

First of all, different implementations of word-space models (such as HAL (Lund et al., 1995), LSA (Landauer and Dumais, 1997), and Random Indexing (Kanerva et al., 2000)) use different kinds of distributional information to produce word spaces. HAL uses word adjacency, LSA uses occurrences in documents, and RI can be used with both of these types of distributional information. Considering that these implementations use different kinds of information to assemble word spaces, it would seem natural to assume that the spaces would contain different kinds of semantic content. Even so, remarkably few studies have investigated how different kinds of distributional information effects the representations. Lavelli et al. (Lavelli et al., 2004) is one of the very few.

To make matters even worse, it is not even obvious what "meaning" means in this context. When we talk about meaning in general discourse, we include a considerable amount of extralinguistic knowledge into the concept of meaning. Part of what I know when I say I know the meaning of "mitt" is what kind of object "mitt" refers to. Such information is arguably not available to word-space models that only consider intralinguistic distributional regularities as data. Although a word-space model might correctly associate "mitt" with "pad" and "glove", it will not be able to reach out into the world and pick out the right kind of object. Thus, "meaning" obviously has a more specific meaning in the context of word-space research, but few — if any — publications further explain what this meaning is. We are left guessing what "meaning" means in word-space research.

Conceptual opaqueness is all too often neglected in favor of experimentalism within the field of computational linguistics. Granted, empirical evidence *should* weigh just as heavy as theoretical arguments, but this is only true if we know what the evidence are evidence *of*. The point is that there can be no evidence unless there is a case. One may seriously question the validity of the research when neither the conceptual nor the evaluative basis are well founded.

The problem with accepting too light-heartedly frail or even ill-advised evaluation methodologies is especially severe

when the experimental models are treated as standard tools that are used to build langauge resources, since any latent flaws in the underlying machinery will inescapably affect the quality of the resource. Consider the not too uncommon case where a word-space model is used to compile a lexical resource, or to solve a retrieval or categorization task: unless we know what kind of information is captured in the word-space model, we will not know what kind of information the lexical resource contains, or why the retrieval or categorization task succeeded or not.

3. Evaluation methodologies in word-space research

In an attempt to taxonomize word-space evaluation practices, we can make a distinction between *direct* and *indirect* evaluation methodologies.

Direct evaluations are concerned with the geometry of the word space, which typically means measuring Euclidean distances between words. The idea is that if word A and word B are closer to each other in the word space than to word C, they are assumed to be more semantically similar to each other than to word C; distance in word space reflects semantic similarity. Of course, there are a number of different measures available for calculating the distance or similarity between objects in an Euclidean vector space.[1] Examples of commonly used measures are the cosine of the angles between the vectors,[2] and different Minkowski metrics.[3] Note that, although these measures *do* produce different similarity scores for a given vector space, they do not change the underlying model.

These geometric measures can be evaluated by comparing them to similarities found in human artefacts such as lexica, priming data, association norms, synonym tests, antonym tests, etc. For artefacts that constitute semantic repositories, such as lexica, priming data and association norms, the evaluation measure is how close the word space resembles the repository — e.g. the fraction of words that occur in both the repository entries and in the word-space neighborhoods. For vocabulary tests, such as synonym and antonym tests, the evaluation measure is performance (normally percentage of correct answers) in solving the test.

Indirect evaluations, on the other hand, are not directly concerned with the geometry of word spaces. Instead, these evaluations *apply* word spaces for various kinds of applications and tasks, the execution of which are normally assumed to require semantic knowledge. Examples include information retrieval and information filtering, word sense disambiguation, text summarization, text categorization, etc.

[1] The difference between distance and similarity measures is that the former produce a low score for similar objects, whereas the latter produce a high score for the same objects: small distance equals large similarity, and conversely. It is trivial to transform a distance measure $dist(x, y)$ into a similarity measure $sim(x, y)$ by e.g. computing $sim(x, y) = \frac{1}{dist(x,y)}$.

[2] $sim_{cos}(\vec{x}, \vec{y}) = \frac{x \cdot y}{|x||y|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}$.

[3] $dist_M(\vec{x}, \vec{y}) = \left(\sum_{i=1}^{n} |x_i - y_i|^N\right)^{\frac{1}{N}}$.

Standardized evaluation metrics exist for most of these tasks; information retrieval and information filtering use precision/recall, word sense disambiguation and text summarization use percentage of correct disambiguations or summarizations, and text categorization normally uses F-scores. There are of course many variations, deviations, and complications related to each of these measures, but the exceptions need not concern us here — the important point is that the evaluation measures for applications used in indirect evaluation of word spaces are well studied and fairly standardized. In fact, some of these applications are even bestowed with entire conference series devoted to evaluation issues. Information retrieval and information filtering has TREC, CLEF and NTCIR; word sense disambiguation has SENSEVAL.

4. Evaluating evaluations

Now, imagine that we have produced a word space. Which evaluation method should we then choose in order to evaluate our word space? Let us consider how we can determine the quality of an evaluation methodology.

One way to accomplish this is to compare the evaluations in terms of reliability and validity. Reliability means that the evaluation metric is consistent, and that it produces roughly the same result every time it is used. As an example, consider an IQ test in which a test subject's results fluctuates between moronic and genius level. Would we say that such a test is useful? Probably not. It is difficult, if not impossible, to draw any conclusions from experiments using a test whose results are inherently inconsistent.

Validity means that the evaluation actually measures what it is supposed to measure. Again, consider the IQ test: does that test really measure intelligence, or is it something else we measure when using it? Without venturing into this highly explosive territory, we can at least maintain that the validity of the IQ test has been subject to much heated debate.

Preferably, we would like our evaluation schemes to be both reliable and valid. The question is: are they?

Let us start with the concept of reliability. Are the above-mentioned evaluation methodologies reliable? Do they produce consistent results? The answer is, for the most part, yes. As a general rule, we can say that the simpler the test, the more reliable it is. This means that tests that involve more parameters, such as parameterized scoring functions or machine learning algorithms that need to be tuned, are generally harder to reproduce than tests that are inherently non-parametric, such as vocabulary tests. From this persepctive, one could easily get the impression that direct evaluations might be more reliable than indirect ones. However, one should keep in mind that the raison d'être of the above-mentioned conference series, and the existence of the many standardized test collections, is precisely to guarantee the reliability of the evaluations.

Note that we may assess the reliability of a test without necessarily knowing what the test is supposed to measure. As long as a test produces consistent results when repeated, we know that it is reliable, regardless of what it actually measures. However, this does not apply to the concept of validity. On the contrary: *we need to know what it is we are*

supposed to measure in order to assess the validity of a test. Thus, the question is: what is it we want to measure?

5. What do we want to evaluate?

Assembling a word space is a parameter-ridden task, which involves a number of non-arbitrary design decisions that effects the resulting word space, including:

- The type of distributional information used to construct the space.
- The weighting and normalization schemes for the distributional statistics.
- The dimensionality reduction technique used for compacting the space.
- The similarity metric used for computing vector similarities.

Out of these four parameters, the first one has received incomparably the least attention, as noted in Section 2 above. The other parameters have been more thoroughly studied. Examples include Nakov et al. (Nakov et al., 2001), who studied the effects of using different weighting schemes in LSA; Bingham and Mannila (Bingham and Mannila, 2001), who investigated the effects of using different dimensionality reduction techniques; and Weeds et al. (Weeds et al., 2004), who studied the effects of using different similarity measures for computing distributional similarity.

Parameter optimization can be both interesting and important in its own right. This is especially true when a word-space model is being fine-tuned for some particular application. If we want to build a text categorization system that uses word-space representations, we want to make sure the word space is as good as it can be with regards to its ability to produce representations for a categorization algorithm. By the same token, if we want to build a word-sense disambiguation system that uses word-space representations, we want to make sure the word space is as good as it can be with regards to its ability to discriminate between different senses of a word. One might conjecture that it would not be the same parameter settings that are optimal for these two tasks, and so it is important to optimize the parameters for one particular task.

It is arguable that the reliability of evaluations is more important than their validity for studies like these that are concerned with the effects of different parameters. What is important in these evaluations is optimizing the performance of a word space in some particular task with regards to some particular parameter. What is *not* important is whether the evaluation really measures meaning. The focus is on comparative analysis of different parameter settings, rather than evaluating the semantic properties of the word space.

6. Measuring meaning

How could we then proceed if our prime focus of interest is the semantic properties of the word space rather than its performance in some particular application? Imagine that we want to measure how good a model of meaning a given word space is. How can we accomplish this? Merely measuring the performance of the word-space representations in some particular application does not tell us very much about how good a representation of meaning it is. Granted, if a word space solves a synonym test with results approaching those of humans, it obviously must contain a fair amount of information about synonyms. By the same token, if it solves an antonym test with good results, it obviously must contain a fair amount of information about antonyms, and if it leads to good information retrieval performance, it obviously must contain a fair amount of... information?

This is where the cookie starts to crumble. Even if a word space leads to good results in a test with an audible semantic foundation, such as a synonym or antonym test, we only know that the word space contains *that particular type* of semantic relation. We still do not know anything about how it fares with meaning in general. The point here is that our ordinary, and intuitive, concept of meaning involves more than mere synonymy, more than mere antonymy, and more than mere associativity. Using these kinds of semantic tests allows us to measure *aspects* of meaning, but they do not license the general conclusion that a word space is a good model of meaning. Unless what we mean by "meaning" is mere synonymy, mere antonymy, or mere associativity, that is.

The point I want to make here is that if we want to measure meaning, we have better first provide an answer to the question what "meaning" means. Does it mean synonymy? Antonymy? The ability to discriminate between topically different texts? Clearly, the answer to this question determines how we judge the semantic validity of the tests. Thus, the question about the validity of evaluation metrics is brought to its head here. If our prime concern is the semantic properties of the word space, then *we need to know what "meaning" means* before we can determine the validity of an evaluation procedure. Ducking the question about the meaning of "meaning" might seem like a clever tactic that evades a horde of notorious philosophical problems, but we can never hope to defend the semantic validity of an evaluation methodology unless we face up to this hard problem. It is futile to try to measure something without knowing what to measure. We simply will not be able to determine whether a word space is a viable model of meaning until we have explained what we mean by "meaning".

7. Conclusion

In this paper, I have discussed evaluation methodologies for word-space models. I began with a brief introduction to word-space models, before turning to a discussion about the need for critical assessments of evaluation methodologies. I then reviewed the bulk of current word-space evaluation procedures, and discussed how we can determine their quality. I argued that we need to know what it is we want to measure if we are to be able to determine the validity of evaluation methodologies, and I discussed some current foci in word-space research. In the last section, I argued that we need to explain what we mean by "meaning" if we want to devise valid tests that we can use to derive evidence of the semantic pertinence of word-space models.

It is now time to revisit the question I started this paper

with: "how can we determine whether a given word space is a *good* word space?" The answer that emerges out of the discussion in this paper is that this answer depends on what we believe the word space is good *for*. If we believe it can be used to build text representations, then we should use any of the indirect evaluation methodologies that operate on texts, such as information retrieval, information filtering, text categorization, or text summarization. If we believe the word space is a viable model of meaning, then we need to know what we mean by "meaning" before we can start devising pertinent evaluation methodologies.

The lesson in this paper is that we should be cautious with making claims about the semantic nature of the word-space representations based solely on empirical evidence. Unless we can base our claims on a theory of meaning, such interpretations require a considerable leap of faith. I will end this discussion with a plea for a more collective discussion about pertinent evaluation methodologies for meaning models in general, and for word-space models in particular. The information retrieval community's successful evaluation campaigns have proven to be a widely stimulating factor in information retrieval research. Perhaps we should view them as paragons?

If nothing else, this paper demonstrates the need for conference series devoted to discussions about the evaluation of natural language processing techniques and resources.

8. References

Ella Bingham and Heikki Mannila. 2001. Random projection in dimensionality reduction: applications to image and text data. In *Proceedings of the 7th ACM SIGKDD international conference on Knowledge Discovery and Data mining, KDD'01*, pages 245–250.

Scott Deerwester, Susan Dumais, George Furnas, Thomas Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the Society for Information Science*, 41(6):391–407.

Stephen Gallant. 1991a. Context vector representations for document retrieval. In *AAAI Natural Language Text Retrieval Workshop*.

Stephen Gallant. 1991b. A practical approach for representing context and performing word sense disambiguation using neural networks. *Neural Computation*, 3(3):293–309.

Fan Jiang and Michael Littman. 2000. Approximate dimension equalization in vector-based information retrieval. In *Proceedings of the 17th International Conference on Machine Learning, ICML'00*, pages 423–430. Morgan Kaufmann, San Francisco, CA.

Pentti Kanerva, Jan Kristofersson, and Anders Holst. 2000. Random indexing of text samples for latent semantic analysis. In *Proceedings of the 22nd Annual Conference of the Cognitive Science Society, CogSci'00*, page 1036. Erlbaum.

Jussi Karlgren and Magnus Sahlgren. 2001. From words to understanding. In Y. Uesaka, P. Kanerva, and H. Asoh, editors, *Foundations of Real-World Intelligence*, pages 294–308. CSLI Publications.

Thomas Landauer and Susan Dumais. 1997. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104(2):211–240.

Alberto Lavelli, Fabrizio Sebastiani, and Roberto Zanoli. 2004. Distributional term representations: an experimental comparison. In *Proceedings of the 13th ACM conference on Information and knowledge management, CIKM'04*, pages 615–624, New York, NY, USA. ACM Press.

Kevin Lund and Curt Burgess. 1996. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instrumentation, and Computers*, 28:203–208.

Kevin Lund, Curt Burgess, and Ruth Ann Atchley. 1995. Semantic and associative priming in high-dimensional semantic space. In *Proceedings of the 17th Annual Conference of the Cognitive Science Society, CogSci'95*, pages 660–665. Erlbaum.

Scott McDonald and Will Lowe. 1998. Modelling functional priming and the associative boost. In *Proceedings of the 20th Annual Conference of the Cognitive Science Society, CogSci'98*, pages 675–680.

Preslav Nakov, Antonia Popova, and Plamen Mateev. 2001. Weight functions impact on lsa performance. In *Proceedings of the EuroConference Recent Advances in Natural Language Processing, RANLP'01*, pages 187–193, Tzigov Chark, Bulgaria.

Magnus Sahlgren and Richard Cöster. 2004. Using bag-of-concepts to improve the performance of support vector machines in text categorization. In *Proceedings of the 20th International Conference on Computational Linguistics, COLING'04*, pages 487–493.

Magnus Sahlgren and Jussi Karlgren. 2005. Automatic bilingual lexicon acquisition using random indexing of parallel corpora. *Journal of Natural Language Engineering*, 11(3):327–341.

Magnus Sahlgren. 2005. An introduction to random indexing. In H.F. Witschel, editor, *Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering, TKE 2005, Copenhagen, Denmark, August 16, 2005*, volume 87 of *TermNet News: Newsletter of International Cooperation in Terminology*.

Hinrich Schütze. 1992. Dimensions of meaning. In *Proceedings of the 1992 ACM/IEEE conference on Supercomputing, Supercomputing'92*, pages 787–796. IEEE Computer Society Press.

Hinrich Schütze. 1993. Word space. In *Proceedings of the 1993 conference on Advances in Neural Information Processing Systems, NIPS'93*, pages 895–902, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Julie Weeds, David Weir, and Diana McCarthy. 2004. Characterising measures of lexical distributional similarity. In *Proceedings of the 20th International Conference of Computational Linguistics, COLING'04*, pages 1015–1021.