Metric independent evaluation of lexical representation

Daichi Mochihashi, Yuji Matsumoto
Graduate School of Information Science, NAIST
{*daiti-m,matsu*}@*is.aist-nara.ac.jp*

Abstract

This paper proposes an evaluation metric for lexical semantic representations. Based solely on the implicit generative assumption of document cohesion, this metric is common to probabilistic and non-probabilistic models, and independent of similarity functions between words.

With this metric, we compared the two state-of-the-art lexical representations in information retrieval, LSI and pLSI to find in fact pLSI produces a better internal semantic cohesion within documents. Other experiment showed that the usually adopted document collocation is not crucial to reproduce the document cohesion itself.

1 Introduction

Most computational studies on lexical semantics suffer from the lack of adequate measures of their representation. This is largely due to the fact that `semantics' itself is not a well-defined concept to handle. To sidestep this problem, NLP tasks implicitly related to the lexical semantics such as text categorization, semantic disambiguation or information retrieval often treat it as a discriminative problem induced indirectly. For example, Joachims (1998) and Cristianini et al. (2001) seek an appropriate semantic kernel for a text categorization, and apply it with SVMs. Widdows and Dorow(2002) evaluates his result with a fitness to the existent data of WordNet (Fellbaum, 1998), which is not of course an ultimate goal of representation.

A simple, direct measure might be through a perplexity, possibly unigram: but this is not sufficient. Although probabilistic models which have a generative feature as pLSI[1] (Hofmann, 1999) and LDA can generate perplexities in a legitimate way, this is not a requisite, for example widely used LSI (Deerwester et al., 1990) is neither probabilistic nor generative. In addition, with graph-theoretic representations as Widdows (2002) it is even more difficult to allot a probability and perplexity, therefore all these methods are not easy to compare.

Does this mean that a direct evaluation is impossible? We think No: although `semantics' itself is ambiguous, we can alternatively recourse to the generative assumption of semantic cohesion which those semantic representations implicitly based on their building.

With this consideration, we propose an ordinal measure of lexical representations in section 2, which is independent of specific characteristics with the semantic spaces and provides a common basis of evaluation. In section 3, we used this metric with LSI and pLSI, and show that in fact pLSI provides a better representation of cohesion. From this point of view, we show that document collocations usually carried in text categorization or other tasks is not crucial to represent document cohesions themselves in section 4. In section 5, we conclude and discuss about the nature of representations based on the collocation.

2 An Ordinal-based Cohesion Metric

2.1 Background

Most semantic clustering techniques depend on local semantic cohesion between words. For example, Papadimitriou et al.(1998) describes an implicit generative document model in LSI, where (relevance to styles omitted)

(1) a topic mixture and the length of a document is sampled;

(2) *all* words in a document are sampled iteratively to fill the document.

Mixture of unigrams (Nigam et al., 2000), pLSI also have similar generative models of documents: first, a hid-

[1] pLSI has a weak generative model; but it requires a summation over all pseudo-index of documents and therefore is hard to evaluate.

den topic or topic mixture $\{z_n\}$ is sampled and words $w_1 \ldots w_d$ are generated according to the probability $p(w_i|z_n)$.

All these effectively assume that words in a document concentrate around some `centroid' which represents the average content of that document. This means, in other words, there are few "noises" within a document: by comparison to the centroid, neighboring words are frequently, and distant words are scarcely, generated.

Therefore, it can be thought that there exists a basic assumption of local semantic cohesion with these methods. For any method with an (implicit) generative model, there is little intervention between words in a document, or any assumed unit of generation (Figure 2.1). It goes hand in hand as Firth's famous saying "You shall know the word by the company it keeps" (Firth, 1957). The point is that the appropriate words around the centroid are generated in a *comparative* way; therefore, a local semantic cohesion means words in a document are in any way similar each other around its centroid.

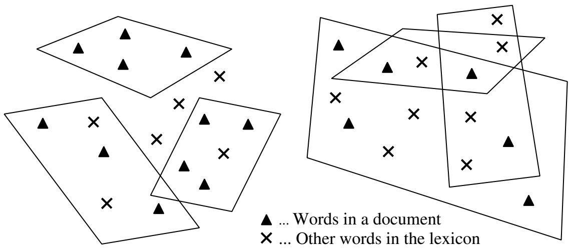


Figure 1: Tessellation of Words and Documents.

Therefore, it seems to be straightforward to evaluate different semantic representations on the basis of their average standard deviation of words within documents. However, this proves to be impossible for some models, since deviation calculation requires an existence of expectation. Although in LSI centroids are easy to calculate, in a probability density representation as pLSI there exists no trivial `expectation of densities' available. In graph-theoretical representation, this becomes even more difficult. Therefore, it is desirable that distance functions are the only requisite to evaluation, which virtually all clusterings would satisfy. Kleinberg (2002) proposes an axiomatic specification of clusterings, where only distances and their related conditions are required. In his three axioms, scale-invariance, richness and consistency, our comparative method especially satisfies consistency, which means absolute coordinates of tessellation is irrelevant to the cohesion results.

In the next subsection, we introduce one of such an ordinal measure of cohesion.

2.2 An Ordinal Cohesion Metric

As stated above, goodness of lexical representation is measured, by definition, through reciprocal local cohesions within a unit of generation (usually a document). This means that, for a word w in a document d, words $\{w_1 \ldots w_d\}$ in d are more likely to be similar than all other words in the lexicon. This comparison does not necessitate specific characteristics of ambient spaces, and diagonal to similarity functions and their scales. Therefore, we can give the following definition of the cohesion of semantic representations.

> Let the collection of all words w be the lexicon L, and a set of its subsets s be S ($S \subset 2^L$.) Let the distance function between words $\delta : L \times L \to \mathbf{R}$. With each s, for each pair of words $(w, w') \in s \times s$, let $r(w, w')$ be the rank of $\delta(w, w')$ within $\{\delta(w, v)|v \in L\}$.
> Then, the average rank of the distance within s is $\bar{r}(s) = 1/(|s|(|s| - 1)) \sum_{(w,w')\in s\times s} r(w, w')$. The average rank through all the structures $s \in S$ gives the overall cohesion, $\bar{r}(S) = 1/|S| \sum_{s\in S} \bar{r}(s)$.

Usually we set the structure (subset of the lexicon) s to be a document, but in principle s can be attributed to any unit of cohesion over which the representation has been built, such as sentences or paragraphs. Through the use of ordinal rank of similarity, this cohesion means comparative proximity within s over other words in the lexicon. Since $r(w, w')$ and therefore its average $\bar{r}(s)$ and $\bar{r}(S)$ varies between $[1, |L| - 1]$, in the maximum cohesion $\bar{r}(S)$ tends to be 1, and the maximum disorder tends to be $|L| - 1$ (average $(|L| - 1)/2$).

In addition to the single unified average $\bar{r}(S)$ over all structures, behavior of cohesion can also be observed by counting the rank distribution of $\bar{r}(s)$ over $s \in S$.

3 Experimental Results

3.1 LSI and pLSI

While pLSI (Hofmann, 1999) is reported to show a better performance of retrieval over LSI, that is an objective measure and it is not an evidence by itself that its internal representation really constitutes an internal semantic cohesion of documents. Therefore, we conducted an experiment to compare the two model based on their generative nature of cohesion, which has so far not been available due to the lack of common ground of comparison.

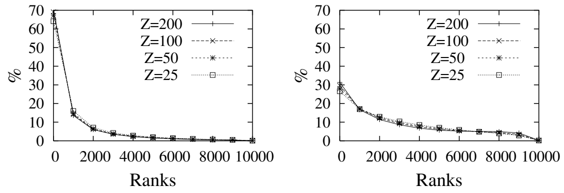
Settings We used Reuters-21578 text categorization data set (Lewis, 1997) with randomly extracted half of ModApte split, which constitutes 4,416 training documents and 1,515 test documents. Standard stop words with Reuters data set and hapax legomena were eliminated and TreeTagger (Schmid, 1994) was applied to

	pLSI	LSI
$\lvert Z\rvert = 200$	12.53%	32.94%
$\lvert Z\rvert = 100$	13.21%	32.89%
$\lvert Z\rvert = 50$	13.91%	33.25%
$\lvert Z\rvert = 25$	14.83%	33.83%

Table 1: Average of Ranks, $\bar{r}(S)$ (%).

produce a 10,101 words of lemmas of lexicon ($|L| = 10,101$). Representations are produced with each of LSI/pLSI, using dimensions 25/50/100/200 each. With LSI, we use a tf.idf as a term weighting [2].

Result Figure 3.1 shows a rank distribution of $\bar{r}(s)$ over each document, and Table 3.1 summarizes it by $\bar{r}(S)$ over all documents. For each model, larger dimensions show a slightly better performance (lover average rank), however pLSI clearly shows a tighter cohesion. $\bar{r}(S)$ in Table 3.2 means, on the average, a word in a document can be predicted from another word as within top 13% of words sorted by similarity for pLSI, and top 33% for LSI.



(a) pLSI (b) LSI

Figure 2: Distribution of Ranks.

3.2 Are Document Collocations Necessary?

Many methods in text categorization or information retrieval adopt a `bag-of-words' assumption of documents, which essentially means a document-wide collocation of words. However, it is questionable that this assumption is adequate, especially for long or structured documents. Beeferman et al. (1997) found that word repetition occurs as an exponential function of interval, which shows flat semantic collocations might be inadequate.

Since our method can detect a document cohesion in a direct way, we measured it with pLSI through document-wide collocation, sentence collocation and some widths of collocational window before and after.

Table 3.2 shows the cohesion $\bar{r}(S)$ and corresponding length of collocational data N. We can see that while small collocational window produces a bad tessellation of words because it relies heavily on syntax, widely used document-wide collocation can be approximated with a sentence collocation or some tens of the length of collocational window. Even some width can produce a better semantic cohesion of documents.

Seeing that there is a large difference between the number of collocations required to build the representations, we can conjecture that a large portion of the document-wide collocation makes little contribution to a document-wide cohesion.

Window	$\bar{r}(S)$	N
1	51.46%	308,672
2	26.34%	584,385
5	12.95%	1,228,421
10	11.37%	1,968,670
20	11.04%	2,904,474
50	10.92%	4,338,420
100	10.98%	5,401,553
Sentence	11.99%	1,804,687
Document	11.08%	6,547,004

Table 2: Collocational window and cohesion.

4 Summary and Discussion

We proposed an evaluation metric of lexical representations directly based on the implicit generative assumption of cohesion. This metric is completely ordinal because it only bases on the relative tessellation of words in a semantic space, therefore independent of absolute scales of perplexity. Although this metric requires no external problem for the evaluation of lexical representations, of course it has some defects. One is that since this metric treats a subset of lexicon appeared in a document, it cannot handle multiple occurrence of words; ideally, words in a document must be represented as a kind of density, because each word has its own weight according to multiplicity or semantic importance. The other is vulnerability to outliers. But we think this effect can be amended through an average over large collections, while of course this metric is not so adequate for a single document.

Besides these problems, experimental result showed that pLSI in fact outperforms LSI in its semantic cohesion within documents.

Experiments with some definitions of collocations showed that document-internal collocation is not always crucial to represent document cohesion, and can be approximated by a sentence collocation or some window functions.

This phenomenon can be regarded as a `semantic n-gram': contrary to the syntactical n-gram language model where larger n would produce more accuracy, this n-gram (although large) smaller n means more accuracy. In other words, in standard n-gram, smaller n implies more data with syntactic ambiguity, while larger n produces more syntactic accuracy at the cost of sparsity; in

[2] Entropy-based term weightings showed a weaker result than tf.idf.

semantic n-gram, larger n implies more data with semantic blurring, while smaller n produces semantic accuracy at the cost of sparsity.

This trade-off must be resolved according to the objectives required. This paper takes the objective as its generative assumption of cohesion, but of course it is not the only criterion. Taking applications into consideration, ideally it is desirable to find a window n adaptively. It will be inevitably task dependent, while this work aimed at a closed evaluation of semantic representations.

References

Nello Cristianini, John Shawe-Taylor, and Huma Lodhi. 2001. Latent Semantic Kernels. In *Proceedings of ICML-01, 18th International Conference on Machine Learning*, pages 66–73.

S. Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by Latent Semantic Analysis. *Journal of the American Society of Information Science*, 41(6):391–407.

Doug Beeferman, Adam Berger and John Lafferty. 1997. A Model of Lexical Attraction and Repulsion. In *Proc. of ACL-EACL '97, Madrid, Spain*, pages 373–380, July.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

J. R. Firth. 1957. A synopsis of linguistic theory, 1930-1955. In *Studies in Linguistic Analysis*, pages 1–32. Oxford: Basil Blackwell.

Thomas Hofmann. 1999. Probabilistic Latent Semantic Indexing. In *Proc. of the 22nd International Conference on Research and Development in Information Retrieval (SIGIR '99)*, pages 50–57.

Thorsten Joachims. 1998. Text categorization with support vector machines: learning with many relevant features. In *Proceedings of ECML-98, 10th European Conference on Machine Learning*, number 1398, pages 137–142.

Jon Kleinberg. 2002. An impossibility theorem for clustering. In *Neural Information Processing Systems 15*.

David D. Lewis. 1997. Reuters-21578 text categorization test collection distribution 1.0. http://www.daviddlewis.com/resources/testcollections/reuters21578/.

Kamal Nigam, Andrew K. McCallum, Sebastian Thrun, and Tom M. Mitchell. 2000. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39(2/3):103–134.

Christos H. Papadimitriou, Hisao Tamaki, Prabhakar Raghavan, and Santosh Vempala. 1998. Latent semantic indexing: A probabilistic analysis. pages 159–168.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, Manchester, UK. unknown.

Dominic Widdows and Beate Dorow. 2002. A Graph Model for Unsupervised Lexical Acquisition. In *COLING 2002: The 19th International Conference on Computational Linguistics Vol.2*, pages 1093–1099.