Optimizing the Dimensionality of Clinical Term Spaces for Improved Diagnosis Coding Support

Aron Henriksson and Martin Hassel

Department of Computer and Systems Sciences (DSV)
Stockholm University, Sweden
{aronhen,xmartin}@dsv.su.se

Abstract. In natural language processing, dimensionality reduction is a common technique to reduce complexity that simultaneously addresses the sparseness property of language. It is also used as a means to capture some latent structure in text, such as the underlying semantics. Dimensionality reduction is an important property of the word space model, not least in random indexing, where the dimensionality is a predefined model parameter. In this paper, we demonstrate the importance of dimensionality optimization and discuss correlations between dimensionality and the size of the vocabulary. This is of particular importance in the clinical domain, where the level of noise in the text leads to a large vocabulary; it may also mitigate the effect of exploding vocabulary sizes when modeling multiword terms as single tokens. A system that automatically assigns diagnosis codes to patient record entries is shown to improve by up to 18 percentage points by manually optimizing the dimensionality.

1 Introduction

Dimensionality reduction is important in limiting complexity when modeling rare events, such as co-occurrences of all words in a vocabulary. In the word space model, reducing the number of dimensions yields the additional benefit of capturing second-order co-occurrences. There is, however, a trade-off between the degree of dimensionality reduction and the ability to model semantics usefully. This trade-off is specific to the dataset—the number of *contexts* and the size of the vocabulary—and, to some extent, the task that the induced term space will be applied to.

When working with noisy clinical text, which typically entails a large vocabulary, it may be especially prohibitive to pursue a dimensionality reduction that is too aggressive. In random indexing, the dimensionality is a predefined parameter of the model; however, there is precious little guidance in the literature on how to optimize—and reason around—the dimensionality in an informed manner. In the current work, we attempt to optimize the dimensionality toward the task of assigning diagnosis codes to free-text patient record entries and reason around the correlation between an optimal dimensionality and dataset-specific features, such as the number of training documents and the size of the vocabulary.

The 4th International Louhi Workshop on Health Document Text Mining and Information Analysis (Louhi 2013), edited by Hanna Suominen.

2 Distributional Semantics

With the increasing availability of large collections of electronic text, empirical distributional semantics has gained in popularity. Such models rely on the observation that words with similar meanings tend to appear in similar contexts [1]. Representing terms as vectors in a high-dimensional vector space that encode contextual co-occurrence information makes semantics computable: spatial proximity between vectors is assumed to indicate the degree of semantic relatedness between terms[2]. There are numerous approaches to producing these context vectors. In many methods they are derived from an initial term-context matrix that contains the (weighted, normalized) frequency with which the terms occur in different contexts[3]. The main problem with using these term-by-context vectors is their dimensionality—equal to the number of contexts (e.g. documents/vocabulary size)—which involves unnecessary computational complexity, in particular since most term-context occurrences are non-events, i.e. most of the cells in the matrix will be zero. The solution is to project the high-dimensional data into a low-dimensional space, while approximately preserving the relative distances between data points. This not only reduces complexity and data sparseness; it has been shown also to improve the accuracy of term-term associations: in this lower-dimensional space, terms no longer have to co-occur directly in the *same* contexts for their vectors to gravitate towards each other; it is sufficient for them to appear in *similar* contexts, i.e. co-occur with the same terms.

This was in fact one of the main motivations behind the development of latent semantic analysis (LSA) [2], which provided an effective solution to the problem of synonymy negatively affecting recall in information retrieval. In LSA, the dimensionality of the initial term-*document* matrix is reduced by an expensive matrix factorization technique called singular value decomposition. Random indexing (RI) [3] is a scalable and efficient alternative in which there is no explicit dimensionality reduction step: it is not needed since there is no initial, high-dimensional term-context matrix to reduce. Instead, pre-reduced d-dimensional[4] context vectors (where $d \ll$ the number of contexts) are constructed incrementally. First, each context (e.g. each document or unique term) is assigned a randomly generated *index vector*, which is high-dimensional, ternary[5] and sparse: a small number (1-2%) of +1s and -1s are randomly distributed; the rest of the elements are set to zero. Ideally, index vectors should be orthogonal; however, in the RI approximation they are—or should be—*nearly* orthogonal[6]. Each unique

[2] This can be estimated by, e.g., taking the cosine similarity between two term vectors.

[3] A context can be defined as a non-overlapping passage of text (a document) or a sliding window of tokens/characters surrounding the target term.

[4] In RI, the dimensionality is a model parameter. A benefit of employing a static dimensionality is scalability. Whether the dimensionality remains appropriate regardless of data size is, we argue, debatable and is preliminarily investigated here.

[5] Allowing negative vector elements ensures that the entire vector space is utilized [4].

[6] There are more nearly orthogonal than truly orthogonal directions in a high-dimensional vector space. Randomly generating sparse vectors within this space will, with a high probability, get us close enough to orthogonality [5].

term is also assigned an initially empty *context vector* of the same dimensionality. The context vectors are then incrementally populated with context information by adding the index vectors of the contexts in which a target term appears.

Although it is generally acknowledged that the dimensionality is an important design decision when constructing semantic spaces [4, 6], there is little guidance on how to choose one appropriately. Often a single, seemingly *magic*, number is chosen with little motivation. Generally, the following—rather vague—guidelines are given: $\mathcal{O}(100)$ for LSA and $\mathcal{O}(1,000)$ for RI [4, 6]. Is this because the exact dimensionality is of little empirical significance, or simply because it is dependent on the dataset and the task? If the latter is the case—and choosing an appropriate dimensionality *is* significant—it seems important to optimize the dimensionality when carrying out experiments that utilize semantic term spaces.

In a few studies the impact of the dimensionality has been investigated empirically. In one study, the effect of the dimensionality of LSA and PLSA (Probabilistic LSA) was studied on a knowledge acquisition task. Dimensionalities ranging from 50 to 300 were tested; however, no regular tendency could be found [7]. In another study using LSA on a term comparison task, it was found that a dimensionality in the 300-500 range was something of an "island of stability", with the best results achieved with 400 [8]. Using RI, there was a study where nine different dimensionalities (500-6,000) were used in a text categorization task. Here the performance hardly changed when the dimensionality exceeded 2,500 and the impact was generally low. The authors were led to conclude that the choice of dimensionality is less important in RI than, for instance, LSA [9].

3 Applying Semantic Spaces in the Clinical Domain

There is a growing interest in the application of distributional semantics to the biomedical domain (see [10] for an overview). Due to the difficulty of obtaining large amounts of clinical data, however, this particular application (sub)-domain has been less explored. There are several complicating factors that need to be considered when working with this type of data, some of which have potential effects on the application of semantic spaces. The noisy nature of clinical text, with frequent misspellings and ad-hoc abbreviations, leads to a large vocabulary, often with concepts having a great number of lexical instantiations. For instance, the pharmaceutical product *noradrenaline* was shown to have approximately sixty different spellings in a set of ICU nursing narratives written in Swedish [11].

One application of semantic spaces in the clinical domain is for the purpose of (semi-)automatic diagnosis coding. A term space is constructed from a corpus of clinical notes and diagnosis codes (ICD-10). Document-level contexts are used, as there is no order dependency between codes and the words in an associated note. The term space is then used to suggest ten codes for each document by allowing the words to vote for distributionally similar codes in an ensemble fashion. In these experiments, a dimensionality of 1,000 is employed [12].

4 Experimental Results

Random indexing is used to construct the semantic spaces with eleven different dimensionalities between 1,000 and 10,000. The data[7] is from the *Stockholm EPR Corpus* [13] and contains lemmatized notes in Swedish. Variants of the dataset are created with different thresholds used for the collocation segmentation [14]: a higher threshold means that stronger statistical associations between constituents are required and fewer collocations will be identified. The collocations are concatenated and treated as single tokens, increasing the number of word types and decreasing the number of tokens per type. Identifying collocations is done to see if modeling multiword terms, rather than only unigrams, may boost results; it will also help to provide clues about the correlation between features of the vocabulary and the optimal dimensionality (Table 1).

Table 1. Data description; the COLL X sets were created with different thresholds.

DATASET	DOCUMENTS	WORD TYPES	TOKENS / TYPE
UNIGRAMS	219k	371,778	51.54
COLL 100	219k	612,422	33.19
COLL 50	219k	699,913	28.83
COLL 0	219k	1,413,735	13.53

Increasing the dimensionality yields major improvements (Table 2), up to 18 percentage points. The biggest improvements are seen when increasing the dimensionality from the 1,000-dimensionality baseline. When increasing the dimensionality beyond 2,000-2,500, the boosts in results begin to level off, although further improvements are achieved with a higher dimensionality: the best results are achieved with a dimensionality of 10,000. A larger improvement is seen with all of the COLL models compared to the UNIGRAMS model, even if the UNIGRAMS model outperforms all three COLL models; however, with a higher dimensionality, the COLL models appear to close in on the UNIGRAMS model.

5 Discussion

There are two dataset-specific features that affect the appropriateness of a given dimensionality: the number of contexts and the size of the vocabulary. In this case, each document is assigned an index vector. The RI approximation assumes the near orthogonality of index vectors, which is dependent on the dimensionality: the lower the dimensionality, the higher the risk of two contexts being assigned similar or identical index vectors[8]. When working with a large number

[7] This research has been approved by the Regional Ethical Review Board in Stockholm (Etikprovningsnämnden i Stockholm), permission number 2012/834-31.

[8] The proportion of non-zero elements is another aspect of this, which is affected by changing the dimensionality while keeping the number of non-zero elements constant.

Table 2. Automatic diagnosis coding results, measured as recall top 10 for exact matches, with clinical term spaces constructed from differently preprocessed datasets (unigrams and three collocation variants) and with different dimensionalities (DIM).

DIM	UNIGRAMS	COLL 100	COLL 50	COLL 0
1000	0.25	0.18	0.19	0.15
1500	0.31	0.26	0.25	0.20
2000	0.34	0.29	0.28	0.24
2500	0.35	0.31	0.30	0.26
3000	0.36	0.33	0.31	0.26
3500	0.37	0.33	0.32	0.28
4000	0.37	0.34	0.34	0.29
4500	0.37	0.33	0.33	0.30
5000	0.38	0.34	0.33	0.30
7500	0.39	0.34	0.33	0.32
10000	0.39	0.36	0.35	0.32
+/- (pp)	+14	+18	+16	+17

of contexts it is important to use a sufficiently high dimensionality. With 200k+ documents in the current experiments, a dimensionality of 1,000 is possibly too low. This is a plausible explanation for the significant boosts in performance observed when increasing the dimensionality. The size of the vocabulary is another important factor. With a low dimensionality, there is less room for context vectors to be far apart [6]. A large vocabulary may not necessarily indicate a large number of concepts—as we saw with the *noradrenaline* example—but it can arguably serve as a crude indicator of that. For instance, the collocation segmentation of the data represents an attempt to identify multiword terms and thus meaningful concepts to model in addition to the (constituent) unigrams. For these to be modeled as clearly distinct concepts, it is critical that the dimensionality is sufficiently large. This is of particular concern when using a wide context definition, as there will be more co-occurrence events, resulting in more similar context vectors. The COLL models are also affected by the fewer tokens per type, which means that their semantic representation will be less statistically well-grounded. The fact that all COLL models are outperformed by the UNIGRAMS model could, however, also be due to poor collocations. Moreover, when working with clinical text, the vocabulary size is typically larger compared to many other domains. This may help to explain why increasing the dimensionality yielded such huge boosts in results also for the UNIGRAMS model.

Compared to prior work [9], where optimizing the dimensionality of RI-based models yielded only minor changes, it is now evident that dimensionality optimization can be of the utmost importance, particularly when working with large vocabularies and large document sets. A possible explanation for reaching different conclusions is their much smaller document set (21,578 vs 219k) and significantly smaller vocabulary (8,887 vs 300k+). It should be noted, however, that their results were already much higher, making it more difficult to increase per-

formance. This can also be viewed as demonstrating the particular importance of dimensionality optimization for more difficult tasks (90 classes vs 12,396).

6 Conclusion

Optimizing the dimensionality of semantic term spaces is important and may yield significant boosts in performance, which was demonstrated on the task of automatically assigning diagnosis codes to clinical notes. It is of particular importance when applying such models to the clinical domain, where the size of the vocabulary tends to be large, and when working with large document sets.

References

1. Harris, Z.S.: Distributional structure. *Word*, 10, pp. 146–162 (1954).
2. Deerwester, S., Dumais, S., Furnas, G., Landauer, T, Harshman, R.: Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6), pp. 391–407 (1990).
3. Kanerva, P., Kristofersson, J., Holst, A.: Random indexing of text samples for latent semantic analysis. In Proceedings of CogSci, p. 1036 (2000).
4. Karlgren, J., Holst, A., Sahlgren, M.: Filaments of Meaning in Word Space. In Proceedings of ECIR, LNCS 4956, pp. 531–538 (2008).
5. Kaski, S.: Dimensionality Reduction by Random Mapping: Fast Similarity Computation for Clustering. In Proceedings of IJCNN, pp. 413–418 (1998).
6. Sahlgren, M.: The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces. In PhD thesis Stockholm University, Stockholm, Sweden (2006).
7. Kim, Y-S., Chang, J-H., Zhang, B-T.: An Empirical Study on Dimensionality Optimization in Text Mining for Linguistic Knowledge Acquisition. In Proceedings of PAKDD, LNAI 2637, pp. 111–116 (2003).
8. Bradford, R.B.: An Empirical Study of Required Dimensionality for Large-scale Latent Semantic Indexing Applications. In Proceedings of CIKM, pp. 153–162 (2008).
9. Sahlgren, M., Cöster, R.: Using Bag-of-Concepts to Improve the Performance of Support Vector Machines in Text Categorization. In Proceedings of COLING (2004).
10. Cohen, T., Widdows, D.: Empirical distributional semantics: Methods and biomedical applications. *Journal of Biomedical Informatics*, 42, pp. 390–405 (2009).
11. Allvin, H., Carlsson, E., Dalianis, H., Danielsson-Ojala, R., Daudaravicius, V., Hassel, M., Kokkinakis, D., Lundgren-Laine, H., Nilsson, G. H., Nytrø, Ø., Salanterä, S., Skeppstedt, M., Suominen, H., Velupillai, S.: Characteristics and Analysis of Finnish and Swedish Clinical Intensive Care Nursing Narratives, In Proceedings of Louhi, pp. 53–60 (2010).
12. Henriksson, A., Hassel, M.: Exploiting Structured Data, Negation Detection and SNOMED CT Terms in a Random Indexing Approach to Clinical Coding. In Proceedings of RANLP Workshop on Biomedical NLP, pp. 11–18 (2011).
13. Dalianis, H., Hassel, M., Velupillai, S.: The Stockholm EPR Corpus: Characteristics and Some Initial Findings. In Proceedings of ISHIMR 2009, pp. 243–249 (2009).
14. Daudaravicius, V.: The Influence of Collocation Segmentation and Top 10 Items to Keyword Assignment Performance. In Proceedings of CICLing, pp. 648–660 (2010).