Part-of-Speech Sense Matrix Model Experiments in the TREC 2004 Robust Track at ICL, PKU

Bing SWEN, Xue-qiang LÜ, Hong-ying ZAN, Qi SU, Zhi-guo LAI, Kun XIANG, Jing-he HU
Institute of Computational Linguistics, Peking University, Beijing 100871, CHINA
http://icl.pku.edu.cn/

1. INTRODUCTION

The Robust Retrieval track is a traditional ad hoc retrieval task with the focus on individual topic effectiveness. This track provides us an opportunity to do experiments on our recently proposed IR model using a word-by-sense matrix document representation, which was called Sense Matrix Model (SMM) [Swen 2003, 2004]. For the first time to extensively test the model, some simpler and easy-to-implement forms of SMM is used for this year's Robust track, where the part-of-speeches of words are treated as the (rough) senses of words. Though the model supports several matrix similarity measures and some advanced data analysis techniques, our initial implementation can only handle sense sets at the scale of a few hundreds of senses. Thus a relatively small part-of-speech tag set is employed and only two different matrix similarity measures used.

In this paper, we describe our model configuration and methods used in the TREC 2004 Robust track. Implementation issues and the submitted runs are also discussed.

2. SENSE MATRIX MODEL

The basic idea of the model is to explicitly introduce both words and senses in the document representation, namely,

$$\text{document } D \Longrightarrow \text{term set} \times \text{sense set}$$
(with association term/sense weights)

A straightforward manner to make use of such combined information in an IR formalism is that we collect the words (feature terms) along with all the senses they actually (or usually) have in the document, and index the document by a term-sense network for retrieval. Such network relationship of words and senses may be further represented by a matrix of weights that represent the association of words and senses, resulting in a matrix representation of documents. A document collection is then represented as a term-sense-document space, in which every sense becomes a term-by-document matrix (and hence the name *SMM*).

Such a matrix-based retrieval model may be regarded as a "sense expansion" of the vector-space model (VSM) [Salton and Lest 1968, Salton 1971]: VSM's document vector of term weights is "expanded" or "split" (distributed) along the sense direction and thus becomes a matrix.

2.1 Measure of Matrix Similarity

There are several possible methods to evaluate the similarity between document matrices. The first one we may think of is a matrix distance defined by an appropriate matrix norm:

$$d(A,B) = \|A - B\|.$$

The Frobenius-norm $\|\cdot\|_F$ and p- (power) norm $\|\cdot\|_p$ are two of the commonly used. Secondly, the concept of "angle" between matrices may also be introduced in correspondence to vector angle. Using a "normalized distance", namely, distance between normalized matrices,

$$d_{\text{norm}}(A,B) = \left\| \frac{A}{\|A\|} - \frac{B}{\|B\|} \right\| \le 2,$$

we may introduce a correct angle

$$\cos\angle_{\text{norm}}(A,B) =_{\text{def}} 1 - \frac{1}{4} d_{\text{norm}}(A,B)^2.$$

When the F-norm is used and the document matrices are vectors, this angle is proportional to the standard vector angle:

$$\cos\angle_{\text{norm}}(\mathbf{q},\mathbf{d}) = \frac{1}{2}(1 + \cos\angle_{\text{VSM}}(\mathbf{q},\mathbf{d})).$$

Thirdly, also note that

$$\cos\angle(A,B) = \frac{\|AB\|}{\|A\| \cdot \|B\|}$$

is the cosine of the "angle" between any two multipliable matrices A and B, based on the *compatibility condition* of matrix norms $\|AB\| \le \|A\| \cdot \|B\|$. In the case of our document matrix, there are two different possible definitions of matrix angles:

$$\cos\angle(D_1,D_2) = \frac{\|D_1 D_2^{\mathrm{T}}\|}{\|D_1\| \cdot \|D_2\|}, \ \cos\angle'(D_1,D_2) = \frac{\|D_1^{\mathrm{T}} D_2\|}{\|D_1\| \cdot \|D_2\|},$$

where $D_1 D_2^{\mathrm{T}}$ is the term-term correlation via senses, and $D_1^{\mathrm{T}} D_2$ is the sense-sense correlation via terms. We tend to use the former since IR is traditionally about the retrieval of documents using terms to match related term sets.

Other possible similarity measures including matrix trace based angles, which also have two possibilities,

$$\cos\angle(D_1,D_2) = \frac{\mathrm{tr} D_1 D_2^{\mathrm{T}}}{\|D_1\| \cdot \|D_2\|}, \ \cos\angle(D_1,D_2) = \frac{\mathrm{tr} D_1 D_2^{\mathrm{T}}}{\|D_1\| \cdot \|D_2\|}.$$

2.2 Part-of-Speeches as Senses

Some simpler and straightforward cases of SMM include POS SMM, where the sense dimensions are the part-of-speeches of index terms. It corresponds to splitting the VSM term weights into the weights of a term's part-of-speeches.

When the input text is POS tagged, there are 2 ways to determine the matrix elements. The simple one it to index each "Word/POS" pair as a VSM term, but record the matrix correspondence (otherwise it would result in a "POS VSM" with more restricted terms). The standard VSM term weightings are directly applicable to these tagged terms.

The other way is to split the VSM weights with POS distributions, with the document matrix form

$$D = \begin{pmatrix} p_{1,1} w_1 & \cdots & p_{1,m} w_1 \\ p_{2,1} w_2 & \cdots & p_{2,m} w_2 \\ \vdots & \cdots & \vdots \\ p_{N,1} w_N & \cdots & p_{N,m} w_N \end{pmatrix},$$

where m is the number of part-of-speeches adopted and N is the term number in the collection. The $p_{i,j}$ parameter may be estimated to be the frequency of the jth POS of word i in document D. The advantage of this method is that for simple applications, the $\{ p_{i,j} \}$ parameters may be set to the POS probability distribution of words in the collection be considered (instead of being computed for each document).

Since current POS tagging has succeeded considerably, the weighting of POS SMM may be expected to be effective.

It is easy to prove that POS SMMs with angle similarity measures based on normalized distances (in the F-norm) or sense-sense correlation matrix trace are equivalent to the POS VSM, which can be regarded as a "flattened SMM". Thus we may directly use the POS VSM to test these SMM instances. This greatly facilitates the design of our experimental system.

3. SYSTEM DESCRIPTION

We implement the features of SMM as extensions to the SMART-11.0 system (Salton and Lesk 1965, Salton 1971, SMART 1992), which in design provides a flexible architecture for adding new IR features and conducting experiments. The space-time efficiency is not yet optimal for our cases. Due to some inherent constraints the experiments are limited to rather small sense sets, and this is the main reason for us to only test the POS SMM cases.

The system has a new preparser (added to SMART) that checks every input term. If it is a tagged word (a pair of "Word/Tag", with Tag being not limited to POS or any other tags), then the preparser adds the word, the tag and "word/tag" as well to the indexing dictionary, recording a matrix row number for the word and a column number for the tag, and a (row, column) pair for word/tag.

If the input is a "plain" word, then the preparser first searches a word/sense-list dictionary, with each text line being a form like

```
bank    106227059/20/9;106800223/14/6;106739355/2/2;201093881/1/1;106250735/1/1;201599940/0/0; \
        201599852/0/0; 201579642/0/1;201393302/0/0; 200841124/0/0;200464775/0/2;109626760/0/0;  \
        109616845/0/0;106800468/0/0;103277560/0/0;102247680/0/0;100109955/0/0
```

where the WordNet sense number, the frequency and the document frequency of each sense of the word "bank" is listed (these t.f and d.f. valuses are obtained from a Brown corpus that comes up with the WordNet). If the word is in the sense dictionary, then for each sense a "word/sense" term is constructed for the later indexing process (some simple sparse data handling methods are introduced, with configurable parameters). If the input word is a new word, then a "word\word" term is used for the matrix element indexing (with '\' replacing '/' for such case). In either case, the (row, column) pair is recorded for each constructed index term.

Matrix computation is straightforward when the matrix elements are recorded this way, ensuring an efficient process of similarity evaluation. On the other hand, SMM with similarity of sense-sense correlation matrix norm or matrix trace is hard to be efficient in the retrieval process, since almost every document matrix shares a large common sense columns with others. Sequential search is implemented in the system, but is not applicable to the TREC data set.

The compromise made is that we use the short <title> field of a topic as keywords in the search of relevant documents via the inverted index. This reduces the number of documents for similarity computation and more importantly, reduces the complexity of retrieval to make it possible to process the large TREC data set, though relevant documents may be overlooked.

The POS SMM experiments used a C++ implemented Brill tagger to tag both the documents and the topics, which is a "transformation-based error-driven learning" POS tagger, with 48 part-of-speeches and a precision of 97.2% on the UPenn WSJ corpus. The tagging was quite efficient in terms of time (less than two day on a Linux PC workstation of a 700MHz CPU and 256MB memory).

4. SUBMITTED RUNS

Using the above implementation of the model, we submitted nine runs for this year's TREC Robust track, described briefly as follows.

Submitted Runs	
icl04pos2t	Indexing only noun and verb words, using the title field, with normalized matrix distance
icl04pos2d	Indexing only noun and verb words, using the description field, with normalized matrix distance
icl04pos2td	Indexing only noun and verb words, using the title and description fields, with normalized matrix distance
icl04pos2f	Indexing only noun and verb words, using the title, description and narrative fields, with normalized matrix distance
icl04pos7td	Indexing the words of 7 merged POS tags from the 48 original tags, using the title and description fields, with normalized matrix distance
icl04pos7f	Indexing the words of 7 merged POS tags from the 48 original tags, using the title, description and narrative fields, with normalized matrix distance
icl04pos7tap	Indexing the words of 7 merged POS tags from the 48 original tags, using the title field, with sense-sense correlation product matrix norm similarity
icl04pos7tat	Indexing the words of 7 merged POS tags from the 48 original tags, using the title field, with sense-sense correlation matrix trace similarity
icl04pos48f	Indexing the words of the 48 original POS tags, using the title, description and narrative fields, with normalized matrix distance

The submitted runs used the default stopword removal, the 'triestem' stemming (after the POS tagging), and the *lnc-ltc* weighting. (Other options were also experimented with partial results recorded.)

The summary measures over 200 old topics, 49 new topics, 50 difficult topics and all topics are listed as follows.

icl04pos2t	0.1499	0.1639	0.0768	0.1526
icl04pos2d	0.1667	0.2072	0.0826	0.1747
icl04pos2td	0.1829	0.2133	0.0950	0.1889
icl04pos2f	0.2115	0.2347	0.1196	0.2160
icl04pos7td	0.1717	0.2052	0.0871	0.1783
icl04pos7f	0.2014	0.2250	0.1071	0.2060
icl04pos7tap	0.0865	0.0529	0.0346	0.0799
icl04pos7tat	0.1767	0.1786	0.0887	0.1771
icl04pos48f	0.1743	0.2161	0.0872	0.1825

Other different weighting schemes were also experimented. An interesting phenomenon found is that for the *nnn-nnn* weighting, precision increases with the size of POS tags, while for the *atc-atc* and *lnc-ltc* weighting, precision goes the other way. On the other hand, stemming seemed to have reduced (or stabilized) the effects of tagging. The performance of the experimented runs came quite close and was of relatively low values. This may be regarded to indicate that smaller tag sets, though may lead to effective tagging, would contribute less to the model. In the future, we plan to have more investigation on this aspect.

5. TOPIC DIFFICULTY PREDICTION

For the task of predicting topic difficulty, a topic difficulty model based on word sense ambiguity is proposed. The sense ambiguity of a keyword might be related to the difficulty of document search. Hence the total (or average) ambiguity of the words in a topic may be considered a rough measure of the topic difficulty. Such a measure should be further modified by word collocation, word similarity, word distribution (d.f.) in the collection, and the position (role) of the word in the topic.

The model is formulated as follows:

$$\text{easiness}(w) = [\frac{\text{certainty(w)}}{\text{similar(w)}} + \text{collocation(w)}] \cdot \text{doc(w)} \cdot \text{weight(position(w))}$$

$$\text{easiness}(topic) = \frac{\sum_{w \in topic,\ w \notin stopword} \text{easiness}(w)}{\#word \text{ in topic but not in stopwords}}$$

$$= \frac{\sum_{\substack{w \in topic \\ w \notin stopword}} [\frac{\text{certainty}(w)}{\text{similar}(w)} + \text{collocation}(w)] \cdot \text{doc}(w) \cdot \text{weight(position}(w))}{\#word \text{ in topic but not in stopwords}}$$

Where certainty(w) stands for the certainty (being unambiguous) of word w, collocation(w) is the collocation strength of w, similar(w) relates to effects of similar (replaceable) words, doc(w) is the document distribution of w, and weight(position(w)) is the effect of w at a position/role of title, description or positive, negative.

In the experiments, a simplified version of the model is used. It uses a sense distribution dictionary constructed from WordNet and a sense-tagged Brown corpus to estimate the difficulty of a single word. Then the topic difficulty is measured by the average easiness of words in the topic. Some other variants were also attempted but not included in the submitted runs. The Kendall correlation between predicted and actual difficulty is as follows.

icl04pos2t	0.174
icl04pos2d	0.169
icl04pos2td	0.187
icl04pos2f	0.170
icl04pos7td	0.172
icl04pos7f	0.175
icl04pos7tap	0.176
icl04pos7tat	0.124
icl04pos48f	0.198

Due to time limitation, the word difficulty measure and the topic difficulty prediction had not been used in the retrieval process. We think this is an interesting issue and worth extensive investigation in the future.

6. Conclusion

In our first TREC experiments, we used a relatively small tag set, namely the part-of-speech tags of the Brill tagger, to evaluate the sense-matrix model. The effectiveness of the POS SMM seems to be less obvious or marginal. We think that this may indicate small, highly merged sense sets such as part-of-speeches have insignificant contribution to the model, or other new matrix similarity measures effective for small tag sets need to be investigated. Another practical issue led us to use smaller tag sets is that our current SMART-11.0 based system has inherent constraints on the tag-set size. The inverted index must be within the file size of 2GB (addressable with a signed 32-bit integer), which prevents us from using a more reasonable sense set, such as one that is designed to be tailored specifically according to the TREC data set from the WordNet sense set.

We have continuing research work on these issues, and plan to participate more TREC evaluations using SMM with better sense sets as well as better similarity measures.

7. REFERENCES

[1] Ide, N. and J. Véronis. 1998. Introduction to the Special Issue on Word Sense Disambiguation: The Start of the Art. Computational Linguistics, Vol24 No1.

[2] Krovetz, R. and W. B. Croft. 1992. Lexical Ambiguity and Information Retrieval. ACM Transactions on Information Retrieval Systems, Vol. 10(2), 115–141.

[3] Miller, G. 1990. Wordnet: an On-line Lexical Database. In Special Issue: International Journal of Lexicography Vol. 3(4). 235 – 312.

[4] Salton, G. 1971. The SMART retrieval system – Experiments in automatic document processing. Prentice Hall Inc., Englewood Cliffs, NJ.

[5] Salton, G. and M. E. Lesk. 1965. The SMART automatic document retrieval system – an illustration. Communication of the ACM, 8(6): 391-398, June 1965.
[6] Salton, G. and M. E. Lesk. 1968. Computer evaluation of indexing and text processing. In Journal of the ACM, volume 15(1), 8–36, January.
[7] SMART version 11. 1992. Available via anonymous ftp from ftp.cs.cornell.edu
[8] Stokoe, C. et al. 2003. Word Sense Disambiguation in Information Retrieval Revisited. In The 26th ACM-SIGIR Conference on Research and Development in Information Retrieval (SIGIR'03).
[9] Sussna, M. 1993. Word Sense Disambiguation for Free-Text Indexing Using a Massive Semantic Network. In Proceedings of the 2nd International Conference on Information and Knowledge Management (CIKM), 67 – 74, Washington, DC.
[10] Swen, Bing (SUN Bin). 2003. Relative Information and a Sense Matrix Model for IR. Technical Report TR-003, ICL, Peking Univ., Nov 2003. (available at http://icl.pku.edu.cn/icl_tr/)
[11] Swen, Bing (SUN Bin). 2004. Sense Matrix Model and Discrete Cosine Transform. In Proceedings of AIRS 2004 (the first Asia Information Retrieval Symposium). Oct 18-20, Beijing, CHINA; LNCS AIRS Proceedings, Springer Verlag, 2004.
[12] Voorhees, E. M. 1993. Using WordNet to Disambiguate Word Sense for Text Retrieval. In Proceedings of the 16th International ACM SIGIR Conference, 171–180, Pittsburgh, PA.