Near-synonym Lexical Choice in Latent Semantic Space

Tong Wang
Department of Computer Science
University of Toronto
tong@cs.toronto.edu

Graeme Hirst
Department of Computer Science
University of Toronto
gh@cs.toronto.edu

Abstract

We explore the near-synonym lexical choice problem using a novel representation of near-synonyms and their contexts in the latent semantic space. In contrast to traditional latent semantic analysis (LSA), our model is built on the lexical level of co-occurrence, which has been empirically proven to be effective in providing higher dimensional information on the subtle differences among near-synonyms. By employing supervised learning on the latent features, our system achieves an accuracy of 74.5% in a "fill-in-the-blank" task. The improvement over the current state-of-the-art is statistically significant.

We also formalize the notion of *subtlety* through its relation to semantic space dimensionality. Using this formalization and our learning models, several of our intuitions about subtlety, dimensionality, and context are quantified and empirically tested.

1 Introduction

Lexical choice is the process of selecting content words in language generation. Consciously or not, people encounter the task of lexical choice on a daily basis — when speaking, writing, and perhaps even in inner monologues. Its application also extends to various domains of natural language processing, including Natural Language Generation (NLG, Inkpen and Hirst 2006), writers' assistant systems (Inkpen, 2007), and second language (L2) teaching and learning (Ouyang et al., 2009).

In the context of *near-synonymy*, the process of lexical choice becomes profoundly more complicated. This is partly because of the subtle nuances among near-synonyms, which can arguably differ along an infinite number of dimensions. Each dimension of variation carries differences in style, connotation, or even truth conditions into the discourse in question (Cruse, 1986), all making the seemingly intuitive problem of "choosing the right word for the right context" far from trivial even for native speakers of a language. In a widely-adopted "fill-in-the-blank" task, where the goal was to guess missing words (from a set of near-synonyms) in English sentences, two human judges achieved an accuracy of about 80% (Inkpen, 2007). The current state-of-the-art accuracy for an automated system is 69.9% (Islam and Inkpen, 2010).

When the goal is to make plausible or even elegant lexical choices that best suit the *context*, the representation of that context becomes a key issue. We approach this problem in the *latent semantic space*, where transformed local co-occurrence data is capable of implicitly inducing global knowledge (Landauer and Dumais, 1997). A latent semantic space is constructed by reducing the dimensionality of co-occurring linguistic units — typically words and documents as in *Latent Semantic Analysis* (LSA). We refer to this *level of association* (LoA) as *document LoA* hereafter. Although document LoA can benefit topical level classification (e.g., as in document retrieval, Deerwester et al. 1990), it is not necessarily suitable for lexical-level tasks which might require information on a more fine-grained level (Edmonds and Hirst, 2002). Our experimental results show

Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pages 1182–1190,
Beijing, August 2010

noticeable improvement when the co-occurrence matrix is built on a lexical LoA between words within a given context window.

One intuitive explanation for this improvement is that the lexical-level co-occurrence might have helped recover the high-dimensional subtle nuances between near-synonyms. This conjecture is, however, as imprecise as it is intuitive. The notion of *subtlety* has mostly been used qualitatively in the literature to describe the level of difficulty involved in near-synonym lexical choice. Hence, we endeavor to formalize the concept of subtlety computationally by using our observations regarding the relationship between "subtle" concepts and their lexical co-occurrence patterns.

We introduce related work on near-synonymy, lexical choice, and latent semantic space models in the next section. Section 3 elaborates on lexical and contextual representations in latent semantic space. In Section 4, we formulate near-synonym lexical choice as a learning problem and report our system performance. Section 5 formalizes the notion of subtlety and its relation to dimensionality and context. Conclusions and future work are presented in Section 6.

2 Related Work

2.1 Near-Synonymy and Nuances

Near-synonymy is a concept better explained by intuition than by definition — which it does not seem to have in the existing literature. We thus borrow Table 1 from Edmonds and Hirst (2002) to illustrate some basic ideas about near-synonymy. Cruse (1986) compared the notion of *plesionymy* to cognitive synonymy in terms of mutual entailment and semantic traits, which, to the best of our knowledge, is possibly the closest to a textbook account of near-synonymy.

There has been a substantial amount of interest in characterizing the nuances between near-synonyms for a computation-friendly representation of near-synonymy. DiMarco et al. (1993) discovered 38 dimensions for differentiating near-synonyms from dictionary usage notes and categorized them into semantic and stylistic variations. Stede (1993) focused on the latter and further decomposed them into seven scalable subcategories. By organizing near-synonym variations into a tree structure, Inkpen and Hirst (2006) combined stylistic and attitudinal variation into one class parallel to denotational differences. They also incorporated this knowledge of near-synonyms into a knowledge base and demonstrated its application in an NLG system.

Table 1: Examples of near-synonyms and dimension of variations (Edmonds and Hirst, 2002).

Types of variation	***Examples***
Continuous, intermittent	seep:drip
Emphasis	enemy:foe
Denotational, indirect	error:mistake
Denotational, fuzzy	woods:forest
Stylistic, formality	pissed:drunk:inebriated
Stylistic, force	ruin:annihilate
Expressed attitude	skinny:thin:slim:slender
Emotive	daddy:dad:father
Collocational	task:job
Selectional	pass away:die
Sub-categorization	give:donate

2.2 Lexical Choice Evaluation

Due to their symbolic nature, many of the early studies were only able to provide "demo runs" in NLG systems rather than any empirical evaluation. The study of near-synonym lexical choice had remained largely qualitative until a "fill-in-the-blank" (FITB) task was introduced by Edmonds (1997). The task is based on sentences collected from the 1987 *Wall Street Journal* (WSJ) that contain any of a given set of near-synonyms. Each occurrence of the near-synonyms is removed from the sentence to create a "lexical gap", and the goal is to guess which one of the near-synonyms is the missing word. Presuming that the 1987 WSJ authors have made high-quality lexical choices, the FITB test provides a fairly objective benchmark for empirical evaluation for near-synonym lexical choice. The same idea can be applied to virtually any corpus to provide a fair amount of gold-standard data at relatively low cost for lexical choice evaluation.

The FITB task has since been frequently adopted for evaluating the quality of lexical choice systems on a standard dataset of seven near-synonym sets (as shown in Table 2). Edmonds

(1997) constructed a second-order lexical co-occurrence network on a training corpus (the 1989 WSJ). He measured the *word-word distance* using *t-score* inversely weighted by both distance and order of co-occurrence in the network. For a sentence in the test data (generated from the 1987 WSJ), the candidate near-synonym minimizing the sum of its distance from all other words in the sentence (*word-context distance*) was considered the correct answer. Average accuracy on the standard seven near-synonym sets was 55.7%.

Inkpen (2007) modeled word-word distance using *Pointwise Mutual Information* (PMI) approximated by word counts from querying the *Waterloo Multitext System* (Clarke et al., 1998). Word-context distance was the sum of PMI scores between a candidate and its neighboring words within a window-size of 10. An unsupervised model using word-context distance directly achieved an average accuracy of 66.0%, while a supervised method with lexical features added to the word-context distance further increased the accuracy to 69.2%.

Islam and Inkpen (2010) developed a system which completed a test sentence with possible candidates one at a time. The candidate generating the most probable sentence (measured by a 5-gram language model) was proposed as the correct answer. N-gram counts were collected from *Google Web1T Corpus* and smoothed with *missing counts*, yielding an average accuracy of 69.9%.

2.3 Lexical Choice Outside the Near-synonymy Domain

The problem of lexical choice also comes in many flavors outside the near-synonymy domain. Reiter and Sripada (2002) attributed the variation in lexical choice to cognitive and vocabulary differences among individuals. In their meteorology domain data, for example, the term *by evening* was interpreted as *before 00:00* by some forecasters but *before 18:00* by others. They claimed that NLG systems might have to include redundancy in their output to tolerate cognitive differences among individuals.

2.4 Latent Semantic Space Models and LoA

LSA has been widely applied in various fields since its introduction by Landauer and Dumais (1997). In their study, LSA was conducted on *document* LoA on encyclopedic articles and the latent space vectors were used for solving TOEFL synonym questions. Rapp (2008) used LSA on *lexical* LoA for the same task and achieved 92.50% in accuracy in contrast to 64.38% given by Landauer and Dumais (1997). This work confirmed our early postulation that document LoA might not be tailored for lexical level tasks, which might require lower LoAs for more fine-grained co-occurrence knowledge. Note, however, that confounding factors might also have led to the difference in performance, since the two studies used different weighting schemes and different corpora for the co-occurrence model[1]. In Section 3.2 we will compare models on the two LoAs in a more controlled setting to show their difference in the lexical choice task.

3 Representing Words and Contexts in Latent Semantic Space

We first formalize the FITB task to facilitate later discussions. A test sentence $t = \{w_1, \ldots, w_{j-1}, s_i, w_{j+1}, \ldots, w_m\}$ contains a near-synonym s_i which belongs to a set of synonyms $S = \{s_1, \ldots, s_n\}, 1 \leq i \leq n$. A FITB test case is created by removing s_i from t, and the *context* (the incomplete sentence) $c = t - \{s_i\}$ is presented to subjects with a set of possible choices S to guess which of the near-synonyms in S is the missing word.

3.1 Constructing the Latent Space Representation

The first step in LSA is to build a *co-occurrence matrix M* between words and documents, which is further decomposed by *Singular Value Decomposition* (SVD) according to the following equation:

$$M_{v \times d} = U_{v \times k} \Sigma_{k \times k} V^T_{k \times d}$$

[1] The former used *Groliers Academic American Encyclopedia* with weights divided by word entropy, while the latter used the *British National Corpus* with weights multiplied by word entropy.

Here, subscripts denote matrix dimensions, U, Σ, and V together create a decomposition of M, v and d are the number of word types and documents, respectively, and k is the number of dimensions for the latent semantic space. A word w is represented by the row in U corresponding to the row for w in M. For a context c, we construct a vector $\mathbf{c}$ of length v with zeros and ones, each corresponding to the presence or absence of a word w_i with respect to c, i.e.,

$$\mathbf{c}_i = \begin{cases} 1 & \text{if } w_i \in c \\ 0 & \text{otherwise} \end{cases}$$

We then take this *lexical space* vector $\mathbf{c}_{v\times 1}$ as a *pseudo-document* and transform it into a *latent semantic space* vector $\hat{c}$:

$$\hat{\mathbf{c}} = \Sigma^{-1} U^T \mathbf{c} \tag{1}$$

An important observation is that this representation is equivalent to a *weighted centroid* of the context word vectors: when $\mathbf{c}$ is multiplied by $\Sigma^{-1}U^T$ in Equation (1), the product is essentially a weighted sum of the rows in U corresponding to the context words. Consequently, simple modifications on the weighting can yield other interesting representations of context. Consider, for example, the weighting vector $\mathbf{w}_{k\times 1} = (\sigma_1, \cdots, \sigma_k)^T$ with

$$\sigma_i = \frac{1}{|2(p_{\text{gap}} - i) - 1|}$$

where p_{gap} is the position of the "gap" in the test sentence. Multiplying $\mathbf{w}$ before Σ^{-1} in Equation (1) is equivalent to giving the centroid gradient-decaying weights with respect to the distance between a context word and the near-synonym. This is a form of a *Hyperspace Analogue to Language* (HAL) model, which is sensitive to word order, in contrast to a *bag-of-words* model.

3.2 Dimensionality and Level of Association

The number of dimensions k is an important choice to make in latent semantic space models. Due to the lack of any principled guideline for doing otherwise, we conducted a brute force grid search for a proper k value for each LoA, on the basis of the performance of the unsupervised model (Section 4.1 below).

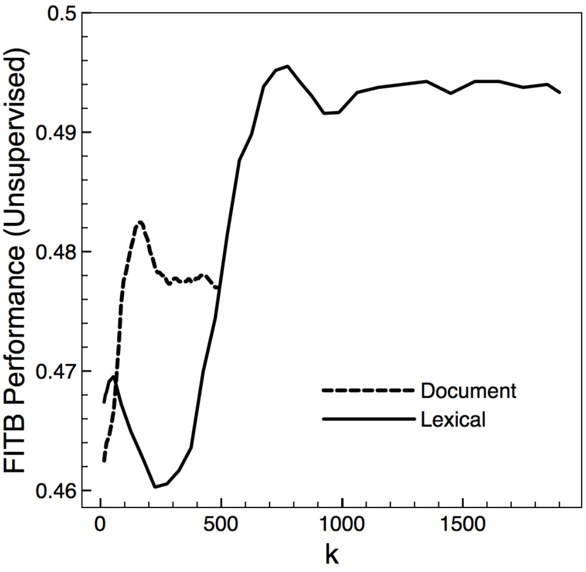

Figure 1: FITB Performance on different LoAs as a function of the latent space dimensionality.

In Figure 1, performance on FITB using this unsupervised model is plotted against k for document and lexical LoAs. Document LoA is very limited in the available number of dimensions[2]; higher dimensional knowledge is simply unavailable from this level of co-occurrence. In contrast, lexical LoA stands out around $k = 550$ and peaks around $k = 700$. Although the advantage of lexical LoA in the unsupervised setting is not significant, later we show that lexical LoA nonetheless makes higher-dimensional information available for other learning methods.

Note that the scale on the y-axis is stretched to magnify the trends. On a zero-to-one scale, the performance of these unsupervised methods is almost indistinguishable, indicating that the unsupervised model is not capable of using the high-dimensional information made available by lexical LoA. We will elaborate on this point in Section 5.2.

[2] The dimensions for document and lexical LoAs on our development corpus are $55,938 \times 500$ and $55,938 \times 55,938$, respectively. The difference is measured between $v \times d$ and $v \times v$ (Section 3.1).

4 Learning in the Latent Semantic Space

4.1 Unsupervised Vector Space Model

When measuring distance between vectors, LSA usually adopts regular vector space model distance functions such as *cosine similarity*. With the context being a centroid of words (Section 3.1), the FITB task then becomes a *k-nearest neighbor* problem in the latent space with $k = 1$ to choose the best near-synonym for the context:

$$s^* = \underset{s_i}{\text{argmax}} \cos(U_{\text{rowId}(\mathbf{v}(s_i),M)}, \hat{\mathbf{c}})$$

where $\mathbf{v}(s_i)$ is the corresponding row for near-synonym s_i in M, and $\text{rowId}(\mathbf{v},M)$ gives the row number of a vector $\mathbf{v}$ in a matrix M containing $\mathbf{v}$ as a row.

In a model with a cosine similarity distance function, it is detrimental to use Σ^{-1} to weight the context centroid $\hat{\mathbf{c}}$. This is because elements in Σ are the singular values of the co-occurrence matrix along its diagonal, and the amplitude of a singular value (intuitively) corresponds to the significance of a dimension in the latent space; when the inverted matrix is used to weight the centroid, it will "misrepresent" the context by giving more weight to less-significantly co-occurring dimensions and thus sabotage performance. We thus use Σ instead of Σ^{-1} in our experiments. As shown in Figure 1, the best unsupervised performance on the standard FITB dataset is 49.6%, achieved on lexical LoA at $k = 800$.

4.2 Supervised Learning on the Latent Semantic Space Features

In traditional latent space models, the latent space vectors have almost invariantly been used in the unsupervised setting discussed above. Although the number of dimensions has been reduced in the latent semantic space, the inter-relations between the high-dimension data points may still be complex and non-linear; such problems lend themselves naturally to supervised learning.

We therefore formulate the near-synonym lexical choice problem as a supervised classification problem with latent semantic space features. For a test sentence in the FITB task, for example, the context is represented as a latent semantic space vector as discussed in Section 3.1, which is then paired with the correct answer (the near-synonym removed from the sentence) to form one training case.

We choose *Support Vector Machines* (SVMs) as our learning algorithm for their widely acclaimed classification performance on many tasks as well as their noticeably better performance on the lexical choice task in our pilot study. Table 2 lists the supervised model performance on the FITB task together with results reported by other related studies. The model is trained on the 1989 WSJ and tested on the 1987 WSJ to ensure maximal comparability with other results. The optimal k value is 415. Context window size[3] around the gap in a test sentence also affects the model performance. In addition to using the words in the original sentence, we also experiment with enlarging the context window to neighboring sentences and shrinking it to a window frame of n words on each side of the gap. Interestingly, when making the lexical choice, the model tends to favor more-local information — a window frame of size 5 gives the best accuracy of 74.5% on the test. Based on *binomial exact test*[4] with a 95% confidence interval, our result outperforms the current state-of-the-art with statistical significance.

5 Formalizing Subtlety in the Latent Semantic Space

In this section, we formalize the notion of subtlety through its relation to dimensionality, and use the formalization to provide empirical support for some of the common intuitions about subtlety and its complexity with respect to dimensionality and size of context.

5.1 Characterizing Subtlety Using Collocating Differentiator of Subtlety

In language generation, subtlety can be viewed as a subordinate semantic trait in a linguistic realiza-

[3] Note that the *context window* in this paragraph is implemented on FITB test cases, which is different from the context size we compare in Section 5.3 for building co-occurrence matrix.

[4] The binomial nature of the outcome of an FITB test case (right or wrong) makes *binomial exact test* a more suitable significance test than the *t-test* used by Inkpen (2007).

Table 2: Supervised performance on the seven standard near-synonym sets in the FITB task. *95% Confidence* based on *Binomial Exact Test*.

	Co-occur.	SVMs	5-gram	SVMs on
Near-synonyms	network	& PMI	language model	latent vectors
	(Edmonds, 1997)	(Inkpen, 2007)	(Islam and Inkpen, 2010)	(Section 4.2)
difficult, hard, tough	47.9%	57.3%	**63.2**%	61.7%
error, mistake, oversight	48.9%	70.8%	78.7%	**82.5**%
job, task, duty	68.9%	**86.7**%	78.2%	82.4%
responsibility, burden, obligation, commitment	45.3%	66.7%	**72.2**%	63.5%
material, stuff, substance	64.6%	71.0%	70.4%	**78.5**%
give, provide, offer	48.6%	56.1%	55.8%	**75.4**%
settle, resolve	65.9%	75.8%	70.8%	**77.9**%
Average	55.7%	69.2%	69.9%	**74.5**%
Data size	29,835	31,116	31,116	30,300
95% confidence	55.1–56.3%	68.7–69.7%	69.3–70.4%	74.0–75.0%

tion of an intention[5]. A key observation regarding subtlety is that it is non-trivial to *characterize* subtle differences between two linguistic units by their collocating linguistic units. More interestingly, the difficulty in such characterization can be approximated by the difficulty in finding a third linguistic unit satisfying the following constraints:

1. The unit must collocate closely with at least one of the two linguistic units under differentiation;
2. The unit must be characteristic of the difference between the pair.

Such approximation is meaningful in that it transforms the abstract *characterization* into a concrete task of finding this third linguistic unit. For example, suppose we want to find out whether the difference between *glass* and *mug* is subtle. The approximation boils the answer down to the difficulty of finding a third word satisfying the two constraints, and we may immediately conclude that the difference between the pair is *not* subtle since it is relatively easy to find *wine* as the qualifying third word, which 1) collocates closely with *glass* and 2) characterizes the difference between the pair by instantiating one of their major differences — the purpose of use. The same reasoning applies to concluding *non-subtlety* for word pairs such as *pen* and *pencil* with *sharpener*, *weather* and *climate* with *forecast*, *watch* and *clock* with *wrist*, etc.

In contrast, for the pair *forest* and *woods*, it might be easy to find words satisfying one but not both constraints. Consequently, the lack of such qualifying words — or at least the relative difficulty for finding one — makes the difference between this pair more subtle than in the previous examples.

We call a linguistic unit satisfying both constraints a *collocating differentiator of subtlety* (CDS). Notably, the second constraint puts an important difference between CDSs and the conventional sense of collocation. On the lexical level, CDSs are not merely words that collocate more with one word in a pair than with the other; they have to be *characteristic of the differences* between the pair. In the example of *forest* and *woods*, one can easily find a word exclusively collocating with one but not the other — such as *national forest* but not **national woods*; however, unlike the CDSs in the previous examples, the word *national* does not characterize any of the differences between the pair in size, primitiveness,

[5] The same principle applies when we replace "generation" with "understanding" and "an intention" with "a cognition".

proximity to civilization, or wildness (Edmonds and Hirst, 2002), and consequently fails to satisfy the second constraint.

5.2 Relating Subtlety to Latent Space Dimensionality[6]

As mentioned in Section 4.1, elements of a latent space vector are in descending order in terms of co-occurrence significance, i.e., the information within the first few dimensions is obtained from more closely collocating linguistic units. From the two constraints in the previous section, it follows that it should be relatively easier to find a CDS for words that can be well distinguished in a lower-dimensional sub-space of the latent semantic space, and the difference among such words should *not* be considered subtle.

We thus claim that co-occurrence-based information capable of characterizing subtle differences must then reside in higher dimensions in the latent space vectors. Furthermore, our intuition on the complexity of subtlety can also be empirically tested by comparing the performance of supervised and unsupervised models at different k values. One of the differences between the two types of models is that supervised models are better at unraveling the convoluted inter-relations between high-dimensional data points. Under this assumption, if we hypothesize that subtlety is a certain form of complex, high-dimensional relation between semantic elements, then the difference in performance between the supervised and unsupervised model should increase as the former recovers subtle information in higher dimensions.

As shown in Figure 2, performance of both models is positively correlated to the number of dimensions in the latent semantic space (with correlation coefficient $\rho = 0.95$ for supervised model and $\rho = 0.81$ for unsupervised model). This suggests that the lexical choice process is indeed “picking up” implicit information about subtlety in the higher dimensions of the latent vectors. Meanwhile, the difference between the performance of the two models correlates strongly to k with $\rho = 0.95$. Significance tests on the “difference of *difference*”[7] between their performances further reveal increasing difference in growth rate of their performance. Significance is witnessed in both the F-test and the paired t-test,[8] indicating that the subtlety-related information in the higher dimensions exhibits complex clustering patterns that are better recognized by SVMs but beyond the capability of the KNN model.

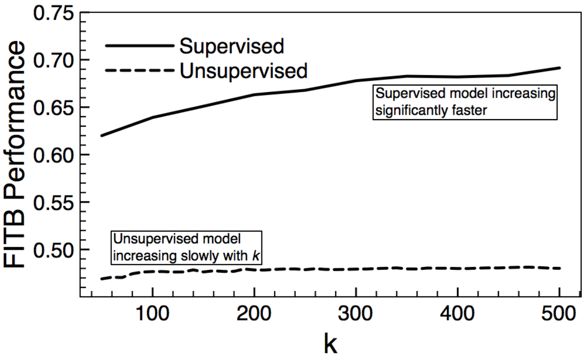

Figure 2: Supervised performance increasing further from unsupervised performance in higher dimensions.

5.3 Subtlety and the Level of Context

Our previous models on lexical LoA associated words within the same sentence to build the co-occurrence matrix. Lexical LoA also allows us to associate words that co-occur in different *levels of context* (LoC) such as paragraphs or documents. This gives an approximate measurement of how much context a lexical LoA model uses for word co-occurrence. Intuitively, by looking at more context, higher LoC models should be better at differentiating more subtle differences.

We compare the performance of models with different LoCs in Figure 3. The sentence LoC model constantly out-performs the paragraph LoC model after $k = 500$, indicating that, by *inter-model comparison*, larger LoC models do not necessarily perform better on higher dimensions. However, there is a noticeable difference in the optimal dimensionality for the model performances. Sentence LoC performance peaks around

[6] In order to keep the test data (1987 *WSJ*) unseen before producing the results in Table 2, models in this section were trained on *The Brown Corpus* and tested on 1988–89 *WSJ*.

[7] The italicized *difference* is used in its mathematical sense as the discrete counterpart of *derivative*.

[8] F-test: $f(1,16) = 9.13, p < 0.01$. Paired t-test: $t(8) = 4.16$ with two-tailed $p = 0.0031$. Both conducted on 10 data points at $k = 50$ to 500 with a step of 50.

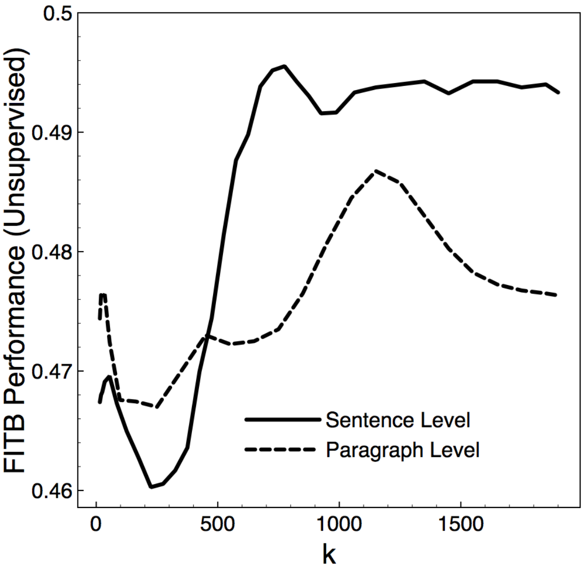

Figure 3: LoC in correlation to latent space dimensionality for optimal model performance.

$k = 700$ — much lower than that of paragraph LoC which is around $k = 1,100$. Such difference may suggest that, by *intra-model comparison*, each model may have its own "comfort zone" for the degree of subtlety it differentiates; models on larger LoC are better at differentiating between more subtle nuances, which is in accordance with our intuition.

One possible explanation for sentence LoC models outperforming paragraph LoC models is that, although the high-dimensional elements are weighed down by Σ due to their insignificance in the latent space, their contribution to the output of distance function is larger in paragraph LoC models because the vectors are much *denser* than that in the sentence LoC model; since the unsupervised method is incapable of recognizing the clustering patterns well in high-dimensional space, the "amplified" subtlety information is eventually taken as noise by the KNN model. An interesting extension to this discussion is to see whether a *supervised* model can consistently perform better on higher LoC in all dimensions.

6 Conclusions and Future Work

We propose a latent semantic space representation of near-synonyms and their contexts, which allows a thorough investigation of several aspects of the near-synonym lexical choice problem. By employing supervised learning on the latent space features, we achieve an accuracy of 74.5% on the "fill-in-the-blank" task, outperforming the current state-of-the-art with statistical significance.

In addition, we formalize the notion of *subtlety* by relating it to the dimensionality of the latent semantic space. Our empirical analysis suggests that subtle differences between near-synonyms reside in higher dimensions in the latent semantic space in complex clustering patterns, and that the degree of subtlety correlates to the level of context for co-occurrence. Both conclusions are consistent with our intuition.

As future work, we will make better use of the easy customization of the context representation to compare HAL and other models with *bag-of-words* models. The correlation between subtlety and dimensionality may lead to many interesting tasks, such as measuring the degree of subtlety for individual near-synonyms or near-synonym sets. With regard to context representation, it is also intriguing to explore other dimensionality reduction methods (such as *Locality Sensitive Hashing* or *Random Indexing*) and to compare them to the SVD-based model.

Acknowledgment

This study is supported by the Natural Sciences and Engineering Research Council of Canada (NSERC). We greatly appreciate the valuable suggestions and feedback from all of our anonymous reviewers and from our colleagues Julian Brooke, Frank Rudzicz, and George Dahl.

References

Charles L. A. Clarke, Gordon Cormack, and Christopher Palmer. An overview of MultiText. *ACM SIGIR Forum*, 32(2):14–15, 1998.

D. A. Cruse. *Lexical Semantics*. Cambridge University Press, 1986.

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

Chrysanne DiMarco, Graeme Hirst, and Manfred Stede. The semantic and stylistic differentiation of synonyms and near-synonyms. *AAAI Spring Symposium on Building Lexicons for Machine Translation*, pages 114–121, 1993.

Philip Edmonds. Choosing the word most typical in context using a lexical co-occurrence network. In *Proceedings of the 35th annual meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages 507–509, 1997.

Philip Edmonds and Graeme Hirst. Near-synonymy and lexical choice. *Computational Linguistics*, 28(2):105–144, 2002.

Diana Inkpen. A statistical model for near-synonym choice. *ACM Transactions on Speech and Language Processing*, 4(1):1–17, 2007.

Diana Inkpen and Graeme Hirst. Building and using a lexical knowledge-base of near-synonym differences. *Computational Linguistics*, 32(2): 223–262, 2006.

Aminul Islam and Diana Inkpen. Near-synonym choice using a 5-gram language model. *Research in Computing Sciences*, 46:41–52, 2010.

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

Shixiao Ouyang, Helena Hong Gao, and Soo Ngee Koh. Developing a computer-facilitated tool for acquiring near-synonyms in Chinese and English. In *Proceedings of the Eighth International Conference on Computational Semantics*, pages 316–319, 2009.

Reinhard Rapp. The automatic generation of thesauri of related words for English, French, German, and Russian. *International Journal of Speech Technology*, 11(3):147–156, 2008.

Ehud Reiter and Somayajulu Sripada. Human variation and lexical choice. *Computational Linguistics*, 28(4):545–553, 2002.

Manfred Stede. Lexical choice criteria in language generation. In *Proceedings of the sixth conference of the European Chapter of the Association for Computational Linguistics*, pages 454–459, 1993.