Identifying Interpersonal Distance using Systemic Features

Maria Herke-Couchman, Casey Whitelaw and **Jon Patrick**
Language Technology Research Group
Capital Markets Co-operative Research Centre
University of Sydney
{*maria, casey, jonpat*}*@it.usyd.edu.au*

Abstract

This paper uses Systemic Functional Linguistic (SFL) theory as a basis for extracting semantic features of documents. We focus on the pronominal and determination system and the role it plays in constructing interpersonal distance. By using a hierarchical system model that represents the author's language choices, it is possible to construct a rich and informative feature representation. Using these systemic features, we report clear separation between registers with different interpersonal distance.

Introduction

This paper explores the categorisation of text based on meaning. Rather than classify on the content matter of a document, we aim to capture elements of the manner in which the document is written. Previous work has looked at extracting other semantic properties of documents. This has included the subjectivity or objectivity of whole texts (Kessler, Nunberg, & Schütze 1997) or individual sentences (Wiebe 1990) (Riloff, Wiebe, & Wilson 2003), and classifying reviews as positive or negative (Turney 2002). Here, we investigate the interpersonal distance, which partially describes the type of relationship established between author and reader. In particular, we use a computational model of Systemic Functional Linguistic theory to extract and represent relevant semantic features.

Much of the prior research has focused on semantic categories of adjectives (Turney 2002) and nouns (Riloff, Wiebe, & Wilson 2003). This paper focuses on the closed class of pronominals and determiners. While the use of these individual words may provide some semantic information, it is through placing them in a system of language choice that patterns of usage may be correlated with interpersonal distance.

We propose preliminary methods for computing aspects of Systemic Functional Linguistics at the lexical level, without dependence on semantic resources or parsers. We show that SFL is well-suited to identifying document-level characteristics of language use.

A Systemic Approach to Interpersonal Distance

Systemic Functional Linguistics (SFL) is a framework for describing and modeling language in functional rather than formal terms. The theory is *functional* in that language is interpreted as a resource for making meaning, and descriptions are based on extensive analyses of written and spoken text (Halliday 1994). The theory is also *systemic* in that it models language as a system of choices (Matthiessen 1995). SFL has been applied in natural language processing in various contexts since the 1960s, but has been used most widely in text generation (Matthiessen & Bateman 1991) (Teich 1995).

Interpersonal distance expresses an aspect of the meaning of the text, and so is located within the semantic statum. As a pattern of meaning, interpersonal distance is realised as a pattern of wording in the lexicogrammar. That is, the meaning is expressed through a pattern of word usage. Similar patterns would be expected to occur in documents that fall into the same register.

A register is a group of texts whose language selections vary from the general language system in similar ways. A register can be characterised by properties of its field, tenor, and mode. Registers are skewings 'of probabilities relative to the general systemic probabilities' (Matthiessen 1993). Register is the instantiation of particular situation types within the system.

While a register groups documents on the basis of the meanings they make, these meanings are realised in the semantics and lexicogrammar of the texts, and so may be analysed on these terms. In particular, registerial differences should be exposed through the patterns of language choice within a system.

Interpersonal Distance

Interpersonal distance refers to the distance between speaker and addressee (Eggins, Wignell, & Martin 1993). Typically, spoken discourse that includes oral and visual contact is representative of minimal interpersonal distance whereas written discourse with no visual, oral or aural contact represents maximal interpersonal distance. However, work on Nigerian emails has indicated that close interpersonal distance might be characteristic of that particular register (Herke-

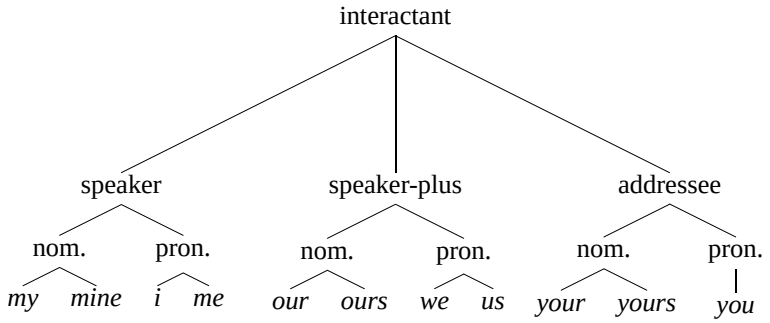

Figure 1: The **interactant** portion of the Pronominal and Determination system

Couchman 2003).

One possible way of measuring the distance between speaker and addressee lexicogrammatically is to explore the language choices made within the pronominal and determination system (Matthiessen 1995).

The Pronominal & Determination System

The Pronominal and Determination system is a language system that includes within it the interpersonal resource for modelling the relationship between the interactants in the dialogue. The system is a closed grammatical system that includes realisations of both interactant (speaker, speaker-plus and addressee) and non-interactant reference items. We use the description of the system given in (Matthiessen 1995).

It is expected that very close interpersonal distance in a text would be characterised by frequent selections from the interactant systems. For example, a text seeking to establish patterns of familiarity between author and reader would show foregrounded patterns of speaker (*I, me, my, mine*) and addressee (*you, your, yours*) usage. Contrastively, a text that is constructing a more formal and distant tenor will typically make little use of the interactant systems but may instead show strong patterns of usage of more generalised alternative meaning systems.

Representing System Networks

For systemic information to be extracted from a document, there must be a suitable computationally-feasible language model. While SFL is a comprehensive and multidimensional linguistic theory, and is not obviously computationally tractable, we can develop a more restricted model that allows us to work with specific systems such as determination.

As is shown by the sample of the system given in Figure 1, this system can intuitively be modelled as a tree. Each internal node represents a subsystem or category: a pattern of possible language choices. Each leaf gives a realisation of its parent system as a word or phrase. A system may contain both lexical realisations and subsystems.

This is an impoverished but still useful view of a system network. Language choice does not always result in a specific word or phrase; an in-depth manual analysis of a text would show that grammatical and lexical units of various sizes contribute to the overall meaning. Further, interaction between systems can result in networks that are not strictly hierarchical, and richer representations will be required to model these processes effectively. The current representation is sufficient to capture language choice for a system such as determination, which is a closed class and fully lexically realised.

Each occurrence of each lexical realisation in the document is counted, and these counts are accumulated upwards through the network. The count at an internal node is the sum of the counts of its sub-categories. This process is no more costly than constructing a feature vector in traditional text classification methods.

In a standard 'bag-of-words' approach, the contribution of a word to a document is given by its relative frequency: how rarely or often that word is used. This implicitly uses a language model in which all words are independent of each other. Crucially, this does not and cannot take into account the *choice* between words, since there is no representation of this choice. Placing words within a system network provides a basis for richer and more informative feature representation.

There are two main advantages gained by adding systemic information for feature representation. Firstly, it allows for categorical features that are based on semantically-related groups of words, at all levels in the network. By collecting aggregate counts, individual variations within a category are ignored. For a given register, it may be the case that important and characteristic language choice occurs at a very fine level, distinguishing between usage of individual words. This word-level information is kept intact, as in a bag-of-words approach. In another register, it may be the usage of a category, such as interactant, that is characteristic. The usage of any words within the category may appear random while maintaining consistent category usage. These higher-level features are not available in a traditional bag-of-words approach, in which these patterns may be lost as noise.

The second and more important difference to traditional feature representation is the representation of language choice. Not only can a system instance calculate the frequency of usage for categories within a system, it can calculate the relative usage within a category. *System contribution* is simply the ratio of sub-category occurrence count to super-category occurrence count, or a normalisation across elements within a category. This gives rise to features such

as ‘interactant usage versus non-interactant usage’. This directly models the fact that in using language, a choice is made. It is a choice not between one word and any other (choosing between unrelated words such as ‘dog’ and ‘elegant’), but between semantic categories within a system. Comparative features such as these can only be used together with a sensible basis for comparison, which is provided here through the use of SFL.

System contribution is not proportional or strongly correlated to term frequency, and the two measures provide useful and complementary information. Term frequency reports the percentage of a document that is made up of a given term. Within a system instance, term frequency can be used to report the term frequency not just of terms but of systems as well. Unlike term frequency, system contribution does not capture how often a system is used, but rather its usage in relation to the other possible choices. In the same way as a register may be characterised by choice, it may also be characterised by frequent usage of a particular system. This gives two complementary representations that may each be useful in characterising semantic features.

Identifying Registers

As discussed, a register is constrained in the types of meanings it is likely to construct. A register may be characterised as establishing a certain interpersonal distance. If the choice within the determination system reflects this semantic position, we should be able to classify documents on this basis.

Not all registers are distinguishable by interpersonal distance. This is but one of many of the semantic properties that characterise documents, such as formality, modality, and evaluation. Note also that the identification of a register is not the same as identifying the *topic* of a document; instances of the ‘newspaper article’ register may have very different content that is presented in the same fashion.

Corpora

We chose corpora that were clearly separated into different registers. From prior manual analysis, it was expected that these registers would have different characteristic interpersonal distance.

Previous work has examined the use of the determination system in so-called ‘Nigerian emails’. These are fraudulent emails in which the author attempts to establish an illegal business relationship (money transfer) with the recipient. One of the most salient characteristics of this register is the way in which the author, despite having no prior relationship with the reader, works to set up a sense of familiarity and trust. These semantic strategies suggest closer interpersonal distance than would usually be expected in the setting up of a legitimate business relationship, particularly since the texts are written rather than spoken. This corpus contained 67 manually collected Nigerian emails.

The Nigerian emails were contrasted with a collection of newspaper articles taken from the standard Reuters text classification corpus. Since many of the newswire texts are very short, only texts with more than one thousand words were kept, resulting in 683 documents. As a result of the context in which they unfold, it was expected that the Reuters newswire texts would make different language choices in order to realise the different meanings they construct. More specifically, it is expected that this register constructs greater interpersonal distance between author and reader.

The third register was taken from the British National Corpus and consists of 195 documents marked as belonging to the ‘spoken / leisure’ category. These are mostly transcriptions of interviews and radio shows covering a wide range of topics. As stated above, the interpersonal distance constructed in spoken text is almost always much closer than that constructed in written texts. Including this corpus allowed us to explore whether the perceived close interpersonal distance in the Nigerian email corpus would be confused with the close interpersonal distance that is typical of spoken texts.

These corpora differ greatly in both field and tenor, and can be separated easily using standard bag-of-words techniques. In using these corpora, we aim not to show improved performance, but to show that the determination system provides sufficient evidence to separate documents on the basis of interpersonal distance. For this to be possible, the words and categories in this system must be used in a regular and learnable fashion, which reflects the semantic positioning of the text.

Features Used

In its entirety, the determination system consists of 109 nodes including 48 lexical realisations. From these, various subsets were used to test the performance and robustness of the system.

all All 109 system and lexis nodes.

lexis The 48 lexical realisations in the system.

system All 61 non-lexical features.

top10 Top 10 features on the basis of information gain

top5 Top 5 features on the basis of information gain

Each set of features was computed once using term frequency (percentage of document) and again using system contribution (percentage of supersystem). Classification was performed using three different machine learners, all commonly used in text classification tasks: a Naive Bayes probabilistic classifier (NB), a decision tree (J48), and a support vector machine (SVM). All implementations are part of the publicly available WEKA machine learning package (Witten & Eibe 1999). As a baseline, we used a standard bag-of-words approach using the top 500 features (ranked by information gain) represented using term frequency. Since the system contribution relies on a structured feature set, no baseline was applicable.

Results

Results from using term frequency and system contribution are shown in Tables 1 and 2 respectively. All of the feature sets and classifiers produced clear separation of the classes, using only features from the determination system. The best result of 99.6% came from an SVM using the system contribution data of either all features or lexical features. It is

	#atts	NB	J48	SVM
all	109	92.8%	98.2%	98.3%
lexis	48	93.8%	98.1%	**98.4**%
system	61	93,9%	98.4%	98.3%
top10	10	96.1%	**98.6%**	97.9%
top5	5	**97.3%**	98.1%	97.8%
baseline	500	98.4%	97.5%	100%

Table 1: classification accuracy using term frequency

	#atts	NB	J48	SVM
all	109	**99.4**%	97.9%	**99.6**%
lexis	48	98.6%	**98.6**%	**99.6**%
system	61	98.6%	98.1%	99.5%
top10	10	98.9%	97.7%	98.6%
top5	5	96.2%	98.1%	98.2%

Table 2: classification accuracy using system contribution

clear from these results that these corpora are separable using features related to interpersonal distance.

Better results were achieved using system contribution than term frequency. By measuring the system choice, rather than system usage, this feature representation highlights the salient aspects of language use. This contrastive description is made possible by placing words in a system network.

In all tests, the Nigerian and Reuters corpora were clearly separated. These registers have markedly different and strongly characteristic interpersonal distance. The spoken corpus exhibited a small amount of confusion with the Nigerian texts, showing evidence that their language is more like spoken than written text.

Feature selection exhibits different effects on the two types of features used. Best performance for system contribution features came from using all features, or only lexical features. Best performance for term frequency features, however, came from using fewer features. Since there is a high degree of correlation between term frequencies within a system network, this can skew results when using classifiers that assume independent features, as Naive Bayes does.

Conclusion

SFL is fundamentally a theory of meaning. As such, language choices can be identified as both formal lexical or grammatical selections as well as in terms of systemic meaning selections. The relationship between these two complementary perspectives is one of abstraction or generalisation; a meaning system is more abstract than the grammar or lexis that realises it (Martin & Rose 2003). This realisation ensures that a meaning phenomenon such as interpersonal distance is characterisable in terms of both systemic choice and lexicogrammatical structure.

In this paper, we have shown that one aspect of the interpersonal distance of a document can be characterised by the use of the determination system. We have further shown that registers that construct variable interpersonal meaning can be separated solely using the features from the Pronominal and Determination system. This can be achieved by modelling SFL at the lexical level without specific external resources.

Interpersonal distance is but one property of the tenor of a document. Similarly, the determination system is but one small part of SFL theory. As our ability to computationally model and extract system networks increases, these systems and their interactions will provide more features by which the semantic properties of a document may be discerned.

References

Eggins, S.; Wignell, P.; and Martin, J. R. 1993. *Register analysis: theory and practice*. London: Pinter. chapter The discourse of history: distancing the recoverable past, 75–109.

Halliday, M. A. K. 1994. *Introduction to Functional Grammar*. Edward Arnold, second edition.

Herke-Couchman, M. A. 2003. Arresting the scams: Using systemic functional theory to solve a hi-tech social problem. In *ASFLA03*.

Kessler, B.; Nunberg, G.; and Schütze, H. 1997. Automatic detection of text genre. In Cohen, P. R., and Wahlster, W., eds., *Proceedings of the Thirty-Fifth Annual Meeting of the ACL and Eigth Conference of the EACL*, 32–38.

Martin, J. R., and Rose, D. 2003. *Working with Discourse: Meaning Beyond the Clause*. London and New York: Continuum.

Matthiessen, C. M. I. M., and Bateman, J. A. 1991. *Text generation and systemic-functional linguistics: experiences from English and Japanese*. London and New York: Frances Pinter Publishers and St. Martin's Press.

Matthiessen, C. M. I. M. 1993. *Register analysis: theory and practice*. London: Pinter. chapter Register in the round: diversity in a unified theory of register, 221–292.

Matthiessen, C. 1995. *Lexico-grammatical cartography: English systems*. International Language Sciences Publishers.

Riloff, E.; Wiebe, J.; and Wilson, T. 2003. Learning subjective nouns using extraction pattern bootstrapping. In *Proceedings of CoNLL-2003*, 25–32. Edmonton, Canada.

Teich, E. 1995. *A Proposal for Dependency in Systemic Functional Grammar – Metasemiosis in Computational Systemic Functional Linguistics*. Ph.D. Dissertation, University of the Saarland and GMD/IPSI, Darmstadt.

Turney, P. D. 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings 40th Annual Meeting of the ACL (ACL'02)*, 417–424.

Wiebe, J. 1990. *Recognizing Subjective Sentences: A Computational Investigation of Narrative Text*. Ph.D. Dissertation, State University of New York at Buffalo.

Witten, I. H., and Eibe, F. 1999. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann.