Automatically Building a Lexicon from Raw Noisy Data in a Closed Domain

Marina Sokolova, Stan Szpakowicz, Vivi Nastase

School of Information Technology and Engineering
University of Ottawa, Ottawa, Canada
{sokolova, vnastase, szpak}@site.uottawa.ca

Abstract

Natural language that people use in electronic communication is far from perfect, due to the narrow channel. This also applies to electronic negotiation. We analyze characteristics of the language data obtained from electronic negotiation. We introduce a novel procedure for extracting and building a lexicon from raw noisy data. The data belong to a closed domain, which allows us to perform domain-dependent word-sense disambiguation. The procedure itself is domain-independent and should work with data from various text collections. We present the results of an application of our procedure to a text corpus collected by an electronic negotiation support system.

1. Introduction

A massive increase of electronic communication has given rise to large, mostly unedited, text collections. When typing replaces face-to-face spoken communication, the language in which people communicate deteriorates. The much narrower channel usually causes inadvertent errors, and there is no time for careful editing. This makes text collections obtained through electronic channels (we call them *Web data*) noisy. We note a large number of spelling and grammatical errors, and the uncontrolled use of informal and slang expressions. The excessive quantity of noise distinguishes Web data from collections of texts communicated through more traditional channels, in particular well edited texts of books, articles and manuals.

Electronic negotiation (e-negotiation) is a rapidly developing domain where people communicate by email or other exchange of text. The management science and Artificial Intelligence communities [1, 6, 11] actively investigate the process and data of electronic negotiation, but nobody seems to have applied natural language processing (NLP) techniques. We employ NLP techniques in a semi-automatic procedure for extracting and building a corpus-based lexicon. This paper presents such a procedure.

We begin by noting that e-negotiation text data share the noise problems of Web data. On the other hand, such texts talk about the well-defined domain of negotiations. These characteristics suggest that the procedure should adjust to noise and benefit from working in a closed domain. Our procedure has both these properties. To investigate its strength, we have applied it to a sample of data from the negotiation support system Inspire [4]. The procedure will help build a preliminary language model for use in e-negotiations. We will use it in a study of the influence of cultural, educational and sociolinguistic background on the process and results of negotiations. While we design the procedure, we keep in mind that Machine Learning (ML) and statistical methods are likely to be used in the next stage of building the model.

When we were planning work on a model of e-negotiation language, we found no NLP study of corpora arising from e-negotiations. So, we must first analyze the characteristics of our sample corpus. A lexicon that reflects such characteristics is the second goal of this study.

In Section 2 we describe the data and discuss the challenges they pose. Section 3 presents a lexicon-construction procedure. Section 4 reports on practical results of the procedure’s application to the Inspire data. Section 5 contains conclusions and suggestions for future work.

2. Data Exploration

We work with a collection of electronic messages exchanged by negotiators that use the Inspire negotiation support system [5]. Negotiation is held between buyers, manufacturers of bicycles, and sellers, manufacturers of bicycle parts. A message may accompany an offer or counteroffer, or can be the only information exchanged at some point during negotiation. Messages have some common features: they are dense, subject-oriented, points of discussion are often accompanied only by salutations and closure, casual talk appears later in negotiation. In casual talk senders exchange personal information, so it contains geographical names (e.g., Marquette, Seward), names of celebrities (e.g., Lord Byron, Celine Dion), names of sport teams, and so on.

We have extracted 14085 messages from the transcripts of 1482 Inspire negotiations. The resulting corpus contains 827209 word tokens and 20990 types. Each negotiation process had new participants, so almost 3000 authors contributed to our corpus. Although they have various educational and cultural background, they share a few characteristics: for most of them English is a second language, they are all enrolled in an MBA program and have all received the same manuals and instructions for negotiation. Although English was suggested as the language for negotiations, some participants used German, Spanish or Russian transliterated in Latin alphabet.

We started the exploration with a manual analysis and gathering of basic statistical information about the original data. Through manual analysis we found that noise originates from the following.

1. Messages with words containing non-letter characters.
2. Text segments in foreign languages, written in ASCII code.
3. Use of foreign words within English text.
4. Use of informal and slang expressions.
5. Spelling errors, missing punctuation and spaces between words, incorrect capitalization.

We consider five matching types of noise. *Noise-corrupted* words are those affected by noise. In Table 1 we give examples of noise-corrupted words in the Inspire data. Here and later in this paper a word's occurrence count in the corpus is shown in brackets, words written in **bold** are spelled correctly, words written in *italic* are noise-corrupted words from the Inspire data. **Deliver**, **negotiate** and **receive** and their word forms are the most often misspelled words among negotiation-related words. In Table 2 we show some of their misspelled versions. We also show spelling versions of the two most often misspelled generic words, **Sincerely** and **Unfortunately**.

Table 1: Noise examples in the Inspire collection.

noise types	2	3	4	5
noise-corrupted words	*offert* (30) *ich* (24)	*niet*(11) *da*(8)	*tu* (32) *yr* (24)	*monday* (32) *deliverytime* (10)

Table 2: Examples of spelling mistakes in the Inspire collection.

Type	Misspelled versions
delivery(4859)	*delievery (21), delevery (11),delivey (8)*
negotiation (2201)	*negociation (152), negotation (24), negotitation (6)*
negotiate (570)	*negociate (64)*
receive (398)	*receive (51), recive (14)*
Sincerely (844)	*Sincerly (41), Sincerelly (7)*
Unfortunately (320)	*Unfortunatly (19), Unfortunatelly (5)*

We have observed that the placement of noise differs.

- noise of type or is concentrated in big chunks throughout the data,
- noise of type , , is spread throughout the data.

The different types of noise also require different elimination approaches:

- noise of type can be fully automatically eliminated from the data,
- noise of type , , , requires manual intervention.

The procedure presented in section 3 incorporates these findings.

To analyze general linguistic behaviour of data we filtered the noise of type out and calculated the number of tokens, types (unigrams), frequencies of unigrams, bigrams and trigrams, *hapax legomena* and *dis legomena* [3]. *N*-gram frequencies will be used by the lexicon-building step of the procedure. Statistical information suggests that the corpus behaves as an unrestricted language corpus [10], though the subjects of text data belong to a closed domain and texts are written by authors with similar post-graduate education. More statistical results and their analysis will appear in a forthcoming paper.

3. Data Extraction and Lexicon Construction

3.1 The Procedure

The procedure presented here constructs a corpus-based lexicon [10]: a mono-lingual lexicon with general syntactic and domain-oriented semantic information. We have built a system containing a vocabulary extraction program, a lemmatizer and a syntactic and semantic information acquisition program. We use an off-the-shelf spell-checker and a general lexical resource with basic syntactic and semantic information. We prefer a general-purpose lexical resource because many negotiators have a restricted English vocabulary, which cannot be identified by a specialized business or computer dictionary. The procedure can be briefly described as follows.

1. Identify and remove data portions not essential in vocabulary extraction.
2. Identify and separate "lexicon-ready" data.
3. Process the remaining data, try to increase the "lexicon-ready" portion.
4. Build a lexicon from the identified data.
5. Process the leftover data.

All Inspire negotiators should use English. We must detect the not uncommon use of other languages, because text segments in foreign languages do not qualify as data for a mono-lingual lexicon. In the first step of the lexicon-building procedure we identify text portions in foreign languages and separate them from portions written in English that only sporadically include foreign words. This task is simpler than language recognition, which is essential for building a multi-lingual lexicon or for translation [10] and requires language recognition algorithms [8]. An obvious way to identify segments in foreign languages is to detect foreign function words, but noise in data makes it highly unreliable. Instead, we manually search the list of types for long, "foreign-looking", words. We then find messages with such words and delete any message not written in English.

In the second step of the procedure we build a list of types from cleaned-up data. We divide the obtained list in two: words present in the lexical resource (*dictionary words*), and words not found there (*non-dictionary words*). At this stage non-dictionary words included inflected forms (**eagerly**, **overlooking**, **ranked**), foreign words (*bien*, *niet*, *Zdravstvuj*), misspelled words (*effictive*, *goodby*, *neighborhood*), words with non-letter characters(*Fant*^{im}*mas*, *won‘’t*), informal (*Hellloooooooooo*, *thnax*, *Thnks*) and slang (*gona* words, *u*), and proper names (**Australia**, **Fumiko**, **Micheal**, **Sahraj**).

The third step reduces the number of types (*dimensionality of data*). This is achieved by increasing the number of dictionary words, each of which can replace several non-dictionary words. The replacement is achieved by spelling correction and then automatic lemmatization. We use a spell-checker to correct the spelling of non-dictionary words; we employ isolated-word error correction [3], because the high volume of noise makes content-dependent error correction unreliable. To select the correct spelling among those suggested by the spell-checker, we check the number of occurrences of each suggested substitution. We choose a substitution with the most occurrences in the data. For example, the Unix tool `ispell` proposes to substitute the word *budgte* with **budge** (21) or **budget** (35); we select **budget**. To extract the largest possible number of dictionary words we repeat their extraction after spelling correction. We run the lemmatizer on non-dictionary and non-corrected words.

The fourth step equips a dictionary word with syntactic and semantic information. Syntactic information consists of part-of-speech information. Semantic information consists of word senses and domain-dependent facts. Semantic information, which can be tuned to the domain, defines *semantic zones*. Zones classify words into several general topics. The e-negotiation data fall into six zones (the last of them comprises closed-category words):

- business in general,
- negotiation processes,
- communication,
- bicycle parts,
- casual talk,
- function words.

The first four zones vary by the domain of the text collection.

The procedure gets dictionary words tagged by semantic category tags from the lexical resource. For each zone tag we find which set of category tags gives the same semantic information. We seek an *automatic* mapping between zone tags and sets of semantic category tags.

We say that a mapping classifies a word correctly if one of its categories corresponds to a zone, within which the word appears in the corpus. Classified words are manually checked using frequencies of bigrams (and trigrams, if necessary). As we plan soon to use ML to classify words from additional information provided by the lexical resource, manual intervention is only temporary. Among all available mappings we choose the mapping that gives the smallest number of misclassified words. Now we can describe the step of the procedure that tags dictionary words with syntactic and semantic information. To obtain syntactic and word sense information we extract part-of-speech (POS) and word sense values from the lexical resource and tag dictionary words with those values. To obtain zone information we extract the values of its category tags in the lexical resource, find the zones corresponding to the categories, and tag the word with the zones.

The last step works with non-dictionary words. We look for personal and geographical names. The search for personal names is automatically done by identifying them in salutations and closures.

3.2 Applicability to Other Data

Though our data exploration and lexicon construction procedure was designed for specific data, we can apply it to Web data gathered from sources other than the Inspire system. For example, a large collection of email texts would be similar to our data. Informal language and use of non-standard symbols such as "smileys" is common in email correspondence. The size of the vocabulary and the amount of noise increase and become more varied when email is written by many senders. Hence, with minimal changes, our procedure can be applied to building a language model of the data obtained from email collections. This could be useful because a corpus-based language model of email communication has not been built yet. Although there is much completed and on-going research on the language in email, we cannot find any references to work with email collections large enough to apply corpus linguistics methods, or with the problem of noise in text data.

4. Empirical Results

As we said in section 3, we tested the lexicon-building procedure on the Inspire collection. After deleting words corrupted by the noise type 1, it contained 20784 types. After deleting messages written in foreign languages, and words containing non-letter characters, the number of types dropped to 14608. In this step of the procedure we could separate dictionary from non-dictionary words. Unprocessed dictionary words were a relatively small part of the Inspire collection: 3255 out of 14608 types. Spelling correction (using Unix's `ispell`) and lemmatization increased the number of dictionary word types to 6512. The total number of types became 9634 because 6120 misspelled types were corrected and replaced by 4061 correct types, which added just 330 new dictionary word types and 1545 non-dictionary types. 1128 inflected form types were replaced by 820 lexemes, which added just 258 new dictionary types. We report the results in Table 3.

Table 3: Dictionary vs non-dictionary words.

Types	Before correction	After correction	After lemmatization	Without names
Dictionary	3255	6354	6512	6512
Non-dictionary	11353	4250	3122	1453

We have extracted syntactic information for dictionary words from a general-purpose lexical resource, Longman Dictionary of Contemporary English (LDOCE) [7]. We report the results of POS tagging in Table 4.

Table 4: Number of part-of-speech tags for the Inspire collection.

Nouns	Verbs	Adjectives	Adverbs	Preposition
3943	2058	1800	672	88

The categories in LDOCE have a hierarchical structure. For example, BUSINESS includes BUSINESS BASICS, which in turn includes ADVERTISING, COMPANIES, BUSINESS MANAGEMENT, MARKETING, OFFICES, TRADE. For five of the six zones defined in section 3, we constructed a mapping into categories through exhaustive search. The function-word zone does not require search.

The following examples illustrate that exhaustive search is necessary, and justify manual intervention into automatic semantic tagging. LDOCE tags the word "margin" only as PUBLISHING, but the bigrams "profit margin", "low margin", "gross margin" show that for Inspire data the word should be negotiation-related. In some cases bigrams do not provide enough information about a word and we have to use trigrams. For example, the LDOCE category tag of "delivery" is BIRTH. This tag does not correspond to the e-negotiation domain. The most frequent bigrams do not provide enough information to tag the word ("the delivery","upon delivery", "delivery time"). Only the 10 th most frequent bigram "delivery payment" and the 15th bigram "price delivery" relate the word to the business zone. On the other hand, the 2th trigram is "payment upon delivery", the 4 th one is "delivery and payment". So, we put the word "delivery" in the negotiation zone.

Table 5 reports the mapping with the smallest number of misclassified words. In the LDOCE column, "A without B" means that we extracted all words, tagged by A but not by B, because words tagged by both do not belong to the corresponding zone. In most cases, for example when A = BUSINESS and B = DAILY LIFE, these words were used as nicknames. Recall from the section 3 that bicycle parts are the topic of negotiation.

Table 5: Correspondence between the Inspire zones and LDOCE categories.

Zone tag	**LDOCE category tags**
Business	Business without daily life, crime and law without biology
Negotiation	General economics, general sports, birth, death, publishing
Communication	Data processing and computing
Topic of negotiation	General transport, general engineering, general industry, bicycles, cars (all of them without biology)
Casual talk	Daily life without sports, general society, politics, religion

5. Conclusion and Future Work

In this paper we have discussed exploration of raw noisy data. We have presented a procedure of extracting and building a lexicon from such data in a closed domain. The procedure is adjustable to the nature of noise in the data and benefits if the domain is closed. Practical results were obtained by applying the procedure to a sample of Inspire data. As the main outcome, we constructed a syntactic and semantic lexicon. We have shown that due to the properties of the data acquired from electronic communication and the domain-independence of the procedure, this procedure is applicable to similar noisy data. The most obvious application would be to an email corpus.

Our immediate future work will concentrate on further development of the procedure through implementation of new spelling correction procedures, design and implementation of lexicon tuning, and application of ML techniques to meaning acquisition.

Acknowledgment

Support for this work comes from the Natural Sciences and Engineering Research Council of Canada, and from the Social Sciences and Humanities Research Council of Canada.

References

[1] S. Jekat, A. Klein, E. Maier, I. Maleck, M. Mast, J. Quantz. “Dialogue Acts in VERBMOBIL”. *Verbmobil-Report 65* (1995). verbmobil.dfki.de/dialog/publications/

[2] K. Jokinen, T. Hurtig, K. Hynna, K. Kanto, M. Kaipainen and A. Kerminen. “Self-Organizing Dialogue Management”. *Proc 2nd Workshop on Neural Networks and Natural Language Processing (NLPRS)*, Tokyo, Japan (2001).

[3] D. Jurafsky and J. H. Martin. *Speech and Language Processing*. Prentice Hall (2000).

[4] G. E. Kersten. “The Science and Engineering of E-negotiation: An Introduction”. InterNeg Report 02/03 (2003). interneg.org/interneg/research/papers/

[5] G. E. Kersten and G. Zhang. “Mining Inspire Data for the Determinants of Successful Internet Negotiations”. InterNeg Report 04/01 (2001). To appear in *European J. of Operational Research*. interneg.org/interneg/research/papers/

[6] S. O. Kimbrough, S. A. Moore. “On Automated Message Processing in Electronic Commerce and Work Support Systems: Speech Act Theory and Expressive Felicity”. *Transactions in Information Systems* **15** (4) (1997) 321-367.

[7] D. Summers (ed). *Longman Dictionary of Contemporary English*, (4th ed.). Pearson Education: Longman (2003).

[8] E. Ludovik, R. Zachnoki, J. Cowie. “Language Recognition for Mono-and Multi-lingual Documents”. *Proc VEXTAL Conference*, Venice, Italy (1999) 209-214.

[9] M. Mast, H. Niemann, E. Noth, E. G. Schukat-Talamazzini. “Automatic Classification of Dialog Acts with Semantic Classification Trees and Polygrams”. *Learning for Natural Language Processing*, Springer, **1040** (1996) 217-229.

[10] T. McEnery, A. Wilson. *Corpus Linguistics*. Edinburg University Press (2001).

[11] C. Quix, M. Schoop , M. Jeusfeld. “Business Data Management for Business-to-Business Electronic Commerce”. *SIGMOD Record*, **31** (1) (2002) 49-54.

[12] M. Schoop. “A Language-Action Approach to Electronic Negotiations”. *Proc 8th International Working Conf on the Language-Action Perspective on Communication Modelling (LAP 2003)* (2003) 143-160.