Generation of Formal and Informal Sentences

Fadi Abu Sheikha and Diana Inkpen

School of Electrical Engineering and Computer Science
University of Ottawa
Ottawa, ON, K1N6N5, Canada

fabus102@uottawa.ca, diana@site.uottawa.ca

Abstract

This paper addresses the task of using natural language generation (NLG) techniques to generate sentences with formal and with informal style. We studied the main characteristics of each style, which helped us to choose parameters that can produce sentences in one of the two styles. We collected some ready-made parallel list of formal and informal words and phrases, from different sources. In addition, we added two more parallel lists: one that contains most of the contractions in English (short forms) and their full forms, and another one that consists in some common abbreviations and their full forms. These parallel lists might help to generate sentences in the preferred style, by changing words or expressions for that style. Our NLG system is built on top of the SimpleNLG package (Gatt and Reiter, 2009). We used templates from which we generated valid English texts with formal or informal style. In order to evaluate the quality of the generated sentences and their level of formality, we used human judges. The evaluation results show that our system can generate formal and informal style successfully, with high accuracy. The main contribution of our work consists in designing a set of parameters that led to good results for the task of generating texts with different formality levels.

1 Introduction

In this paper, we introduce an important technique that takes into account the differences between the formal text style and the informal text style. This technique is automatic text generation that can generate texts that are formal or informal, based on the user preferences.

There are linguistic studies that state that there are different levels of formality (Hayakawa, 1994). We focus on the coarse-grained level, formal and informal style, but finer-grained levels are possible (e.g., informal, less formal, formal, extremely formal).

The motivation for our work is the need for a software tool that helps people to generate formal or informal texts. One of the difficult issues of writing in English is the knowledge of how to adapt to formal or informal situations. Some situations (such as applying for a job) are likely to be formal, whereas others (such as emailing a friend or family member) are more likely to be informal. The real problem when writing is to know what words, phrases, or expressions to use. There are some words, phrases, and expressions that are either formal or informal; for instance, if the wrong word is chosen, then the reader may think we are being either too friendly or too formal.

The paper is organized as follows: in Section 2 we discuss related work; Section 3 addresses the main differences between the formal and the informal style; Section 4 presents the parameters that we used for generation; Section 5 describes our text generation system; results are shown in Section 6; Section 7 concludes the paper and suggests directions of future work.

2 Related Work

In this section, we briefly explain the natural language generation techniques (Reiter and Dale, 2000), the SimpleNLG package (Gatt and Reiter, 2009), and we discuss some of existing NLG

Proceedings of the 13th European Workshop on Natural Language Generation (ENLG), pages 187–193,
Nancy, France, September 2011.

systems that included stylistic variations.

2.1 Natural Language Generation (NLG)

Natural language generation is the process of constructing a natural language text from non-linguistic representation of information in order to meet specified communicative goals (McDonald, 1987). The aim is to build computer systems that automatically produce correct texts in English, and other human languages (Reiter et al., 1995). The main stages and the architecture of a typical NLG system were introduced by Reiter and Dale (2000).

2.2 SimpleNLG Package

The SimpleNLG[1] package (Gatt and Reiter, 2009) can be used to write a program which generates grammatically correct English sentences. It is a library, written in Java, which performs simple and useful tasks that are necessary for natural language generation. The main task that SimpleNLG performs is sentence realisation, which includes orthography, morphology, and simple grammar.

2.3 NLG and SimpleNLG

Following the architecture of Reiter and Dale (2000), the SimpleNLG performs Surface Realisation[2], which is one of the main components of an NLG System. The Surface Realiser does the following tasks:

- Linguistic realisation: this component uses the grammar rules to convert abstract representations of sentences into actual text.
- Structure realisation: converts sentences and paragraphs into mark-up symbols and displays the text.

2.4 Some NLG Systems that Include Style

There are many NLG systems implemented to generate texts for specific purposes. Many of them are commercial systems. For example, the Forecast Generator (FOG) system was designed in 1992 by CoGenTex[3] to generate weather reports in English and French; the inputs of the system were graphical and numerical weather depictions (Goldberg et al., 1994).

We discuss here related work in NLG systems that take into consideration generating text under pragmatic constraints, especially according to style. As far as we are aware, there are only a few researchers who investigated producing text with varied styles.

Hovy (1988, 1990) introduced an NLG system called PAULINE, which is considered as one of the earliest examples of Natural Language Generation systems. Hovy proposed to generate text under pragmatic constraints, including formality. Although small scale, his experiments generated the same text in different styles, to achieve different effects on the reader, and incorporated some pragmatics into language generation. He suggested using different words to generate different styles.

Stamatatos et al. (1997) proposed a system that can generate business letters based on different user requirements, such as style and tone.

Power et al. (2001) proposed the Iconoclast system that allows the users to choose a number of high-level parameters for the text style. These parameters could be sentence length, frequency of passive voice and pronouns, and the use of technical terms. This system allows the user to choose the parameters by manipulating slider bars in a graphical user interface.

Furthermore, Reiter et al. (2003) presented the STOP system that was developed in University of Aberdeen for the British Health Services; it generates tailored letters to help people stop smoking. The STOP system makes the text friendlier by adding more empathy; it also makes the text easier to read for people with poor reading skills.

3 Formal and Informal Language Style

In this section, we explain the main characteristics of informal versus formal style. We also present the parallel lists of words, phrases, and expressions for both styles, which we collected from different sources. The understanding of the main differences between the styles will help to build a system that generates sentences with formal and informal style, by implementing some of these characteristics in our NLG system.

[1] http://www.csd.abdn.ac.uk/~ereiter/simplenlg/

[2] http://www.ling.helsinki.fi/kit/2008s/clt310gen/docs/simplenlg-tutorial-v37.pdf

[3] http://www.cogentex.com/

3.1 Characteristics of Formal versus Informal Style

We briefly explain and summarize the main characteristics of informal style versus formal style, as we found them described in (Dumaine and Healey, 2003; Obrecht, and Ferris, 2005; Akmajian et al., 2001; Park, 2007; Zapata, 2008; Siddiqi, 2008; Redman, 2003; Rob S. et al., 2008; Pavlidis, 2009; Obrecht, 1999). These characteristics are used for building templates to generate sentences based on them. Here, we explain the characteristics of each style and provide examples:

A. Main Characteristics of Informal Style Text:

- It uses personal pronouns and the active voice.
- It uses short simple words and sentences.
- It uses Contractions (e.g., "won't").
- It uses many abbreviations (e.g., "TV").
- It uses many phrasal verbs.
- The words that express rapport and familiarity are often used in speech, such as "brother", "buddy", and "man".
- It uses a subjective style, expressing opinions and feelings.
- It uses vague expressions and colloquial (slang words are accepted in spoken not in written text (e.g., "wanna" = "want to")).

B. Main Characteristics of Formal Style Text:

- It uses impersonal pronouns and often the passive voice.
- It uses complex words and sentence.
- It does not use contractions.
- It does not use many abbreviations.
- It uses appropriate and clear expressions, business, and technical vocabulary.
- It uses politeness words and formulas such as "Please", "Sir".
- It uses an objective style, using facts and references to support an argument.
- It does not use vague expressions and slang words.

3.2 Formal versus Informal lists

We present our parallel lists of informal versus formal words, phrases, and expressions. These lists were collected manually from different sources: the first list is for formal versus informal words and phrases, the second list is for most of the contractions in English, and the third list is for some of the common abbreviations in English. These lists are important parameters for our system of sentence generation.

A. Informal/Formal list of words and phrases

This is a parallel list for informal versus formal words and phrases. We collected this list manually from different sources: (Gillett et al., 2009; Park, 2007; Redman, 2003; Rob et al., 2008). In addition, we obtained a new list that was extracted manually by Brooke et al. (2010) from the dictionary of synonyms *Choose The Right Word* (Hayakawa, 1994). Table 1 shows a sample of this parallel list.

Informal	Formal
about	approximately
anybody	anyone
ask for	request
buy	purchase

Table 1: Examples of formal and informal words from our parallel list

B. Contractions Lists

This is a parallel list for most of the contractions in English (short forms) that represent the informal style versus the full forms of the contractions that represent the formal style. We obtained this list manually from (Redman, 2003; Garner, 2001; Pearl Production, 2005; Woods, 2010). In Table 2, we show a sample of the parallel list of the contractions and their equivalent full forms.

Informal	Formal
aren't	are not
can't	cannot
I'm	I am

Table 2: Examples of contractions versus their equivalent full forms

C. Abbreviation Lists

This is a parallel list for some of the most common abbreviations in English that represent the informal style versus the full forms that corresponds to these abbreviations as used in formal style. However,

there are some abbreviations that are acceptable in formal texts (Obrecht, 1999). We collected this list manually from (Redman, 2003; Gibaldi, 2003; Pearl Production, 2005). Table 3 shows a sample of pairs from the parallel list of the abbreviations and their equivalent full forms.

Informal	Formal
e.g.	for example
etc.	and so on
Feb.	February
Lab	Laboratory

Table 3: Examples of abbreviations and their equivalent full forms

4 Formality Parameters

In this section, we propose the following two main parameters that will be used in constructing formal/informal sentences. We hypothesize that both parameters might help to produce sentences in both styles.

a. Phrase, expression, and word choice (lexical choice): This parameter might help to generate sentences in both styles (Hovy, 1988). We implement this parameter in our system based on the parallel lists (formal/informal words, the contraction list, and the abbreviation list) that we have described in Section 3.
b. Passive/Active voice option: This parameter is based on the characteristics of both styles which we mentioned in Section 3. In addition, it was suggested by Hovy (1988). We added this parameter to our system and we let the system choose a sentence in the passive or the active voice, based on the preferred style.

5 Formal/Informal Sentence Generation

Our system can generate natural language sentences in a formal/informal style with different inputs of subject, verb, and complement (by complement, we mean one or more words including subordinate clauses, as expected in SimpleNLG). Therefore, the user might not worry about choosing any word that he/she is not very familiar with, whether the word is formal or informal, because the system will manage to replace some words with more appropriate words, based on the desired style. In addition, our system might interact with the user directly, or it can be integrated with any system that has the ability to send and receive commands from Java programs.

In the following, we explain the main steps for our system to generate sentences:

a. Ask the user which style is preferred to be generated in the sentence.
b. Ask the user to enter a template that represents the sentence in the form of a subject, a verb, and the rest of the sentence.
c. Ask the user about some syntactic features: the verb tense (present, past, future), progressive (yes, no), perfect (yes, no), and negation (yes, no).
d. The system then checks the verb in the formal/informal parallel list; if it is formal or informal, and the system will find a synonym of the verb in the list, it will replace it based on the preferred style. In addition, if the chosen style is Formal, then the system will choose to generate a sentence in passive voice.
e. After the sentence is constructed, the system will search for any word, phrase, or expression from the formal/informal list, the abbreviations list, and the contractions list, in order to replace it with a synonym, based on the preferred style.
f. Lastly, our system will generate a natural language sentence according to the preferred style, using SimpleNLG for surface realization.

6 Results and Evaluation

Natural language generation is most often evaluated using scores given by human judges (Reiter and Belz, 2009). Our evaluation target was to measure the degree of formality (Formal / Informal) of the generated sentences. We asked two human judges (graduate students in computational linguistics, native speakers of English) to annotate 100 generated sentences as having formal or informal style. Table 4 shows samples[4] of the generated sentences with the

[4] We will make the test set of annotated sentence available, on our website, in case other researchers need them for testing, as well as the three word lists used by our system.

judges' annotations. We estimate the correctness of our system by comparing the original class of the generated sentences (formal/informal) to the annotations of Judge1 and to the annotations of Judge2. We calculated several evaluation measures, to see if our proposed system achieves good quality in producing English sentences in formal and informal style. These measures are the accuracy (correctness) of our system according to each judge, and the precision for each class according to each judge.

Sentence	Actual Class	Judge1 annotate	Judge2 annotate
The plane is going to leave on Jan. 5th.	Informal	Formal	Informal
They were transmuting the raw materials to finished goods.	Formal	Formal	Formal

Table 4: Samples of the generated sentences with the annotations from both judges

		Predicted Class		Precision
		Informal	Formal	
Actual Class	Informal	TP = 45	FN = 0	0.90
	Formal	FP = 5	TN = 50	1.00

Table 5: The results compared to the annotations of Judge1, with the precision for each class

		Predicted Class		Precision
		Informal	Formal	
Actual Class	Informal	TP = 50	FN = 1	1.00
	Formal	FP = 0	TN = 49	0.98

Table 6: The results compared to the annotations of Judge2, with the precision for each class

The results of the annotations show high accuracy for the generated sentences. In Table 5 and Table 6, we show the results according to each of the two judges. The accuracy of our system is 95% according to Judge1 and 99% according to Judge2.

We also calculated the agreement between the two judges, and the kappa statistic that compensated for agreement by chance (Cohen, 1960) (Manning et al., 2008). The agreement between the two judges is 94% and the kappa value is 0.88. This shows a very good agreement for the task.

7 Conclusion and Future Work

In this paper, we have addressed the task of generation of formal and informal texts. The main characteristics of formal and informal style that we identified are success factors for our work, because they helped us to build the parameters that lead to good generation results. In addition, the parallel lists of formal versus informal words and phrases that we collected from different sources were very important in designing our system for the generation formal and informal sentences.

We developed an NLG system that can generate formal and informal sentences. We used template-based NLG techniques in the SimpleNLG package in order to implement our system. We proposed some important parameters that are used in generating formal and informal sentences. We think that these parameters were selected successfully because the evaluation with human judges showed a high accuracy in generating formal and informal sentences. Generating sentences with different formality levels is very useful for various applications (e.g., generating feedback for e-learning games, letters to clients, and other formal or informal documents).

Our future work will be on extracting more formal and informal lists; this should increase the possibility of generating more and more formal/informal sentences, with high accuracy. We will apply different techniques, such as bootstrapping, which can be used in order to extract more lists of words, based on some seed words. We also plan to extend the implementation of our NLG system to cover generating longer texts (e.g., generating several sentences, by adding aggregation, or replacing some nouns with pronouns to avoid repetitions).

References

Akmajian, Adrian, Demers, Richard A., Farmer, Ann K., and Harnish, Robert M. 2001. Linguistics: an introduction to language and communication. 5th Edition, MIT Press, Cambridge (MA), pp. 287-291.

Brooke, Julian, Wang, Tong, and Hirst, Graeme. 2010. Inducing Lexicons of Formality from Corpora. Proceedings. Workshop on Methods for the Automatic Acquisition of Language Resources and their Evaluation Methods, 7th Language Resources and Evaluation Conference, 17-22 May, Valetta, Malta, pp. 605—616.

Cohen, Jacob. 1960. Coefficient of agreement for nominal scales. Educational and Psychological Measurement. VOL. XX, No. 1, pp. 37-46.

Dumaine, Deborah, and Healey, Elisabeth C. 2003. Instant-Answer Guide to Business Writing: An A-Z Source for Today's Business Writer. 2003 Edition, Writers Club Press, Lincoln, pp. 153-156.

Garner, Bryan A. 2001. A Dictionary of Modern Legal Usage. 2nd Edition, pp. xxv, Oxford University Press, US.

Gatt, Albert, and Reiter, Ehud. 2009. SimpleNLG: A realisation engine for practical applications. In Proceedings of the 12th European Workshop on Natural Language Generation (ENLG 2009), 30-31 March, Athens, Greece, pp. 90-93.

Gibaldi, Joseph. 2003. MLA Handbook for Writers of Research Papers. 6th edition, Section 7.4, Modern Language Association of America, (ISBN: 0873529863 / 0-87352-986-3).

Gillett, Andy, Hammond, Angela, and Martala, Mary. 2009. Inside Track to Successful Academic Writing. Pearson Education, ISBN: 978-0273721710.

Goldberg, E., Driedger, N., and Kittredge, R. I. 1994. Using Natural Language Processing to Produce Weather Forecasts. IEEE Expert: Intelligent Systems and Their Applications, 9(2): 45-53.

Hayakawa, S. I., editor. 1994. Choose the Right Word: A Contemporary Guide to Selecting the Precise Word for Every Situation. 2nd Edition, revised by Eugene Ehrlich. HarperCollins Publishers, NY, USA.

Hovy, Eduard H. 1988. Generating Natural Language under Pragmatic Constraints. Lawrence Erlbaum Associates, Hillsdale, NJ, USA, pp. 82 – 87.

Hovy, Eduard H. 1990. Pragmatics and natural language generation. AI. vol. 43, pp. 153–197.

Manning, Christopher D., Raghavan, Prabhakar, and Schütze, Hinrich. 2008. Introduction to Information Retrieval. Cambridge University Press, Chapter 8, pp. 165- 166.

McDonald, David D. 1987. Natural language generation. In Stuart C. Shapiro, editor, Encyclopaedia of Artificial Intelligence, John Wiley and Sons, pp. 642-655.

Obrecht, Fred 1999. Minimum Essentials of English. 2nd Edition, Barron's Educational Series Inc., Los Angeles Pierce College, New York, page 13.

Obrecht, Fred, Ferris, Boak. 2005. How to Prepare for the California State University Writing Proficiency Exams. 3rd Edition, Barron's Educational Series Inc., New York, page 173.

Park, David. 2007. Identifying & using formal & informal vocabulary. IDP Education, the University of Cambridge and the British Council, the Post Publishing Public Co. Ltd.

Pavlidis, Mara. 2009. Target Your Study Skills: Optimise Your Learning. Faculty of Health Sciences. La Trobe University, Victoria, Australia, Chapter 5, page 3.

Pearl Production (Ed). 2005. English Language Arts Skills & Strategies Level 5. Saddleback Publishing, Inc. USA, ISBN 1-56254-839-5.

Power, Richard, Scott, Donia, and Bouayad-Agha, Nadjet. 2003. Generating texts with style. In: Proceedings of the 4th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing'03), pp. 444-452.

Redman, Stuart. 2003. English vocabulary in use: Pre-intermediate & intermediate. 2nd Edition, Cambridge University press, UK.

Reiter, Ehud and Belz, Anja. 2009. An Investigation into the Validity of Some Metrics for Automatically Evaluating Natural Language Generation Systems. Computational Linguistics 25:529–558.

Reiter, Ehud and Dale, Robert. 2000. Building Natural Language Generation Systems (Studies in Natural Language Processing). Cambridge University Press 2000. ISBN 0-521-62036-8.

Reiter, Ehud, Mellish, Chris, and Levine, John. 1995. Automatic Generation of Technical Documentation. Applied Artificial Intelligence 9, pp. 259-287.

Reiter, Ehud, Robertson, Roma, and Osman, Liesl. 2003 Lessons from a Failure: Generating Tailored Smoking Cessation Letters. Artificial Intelligence, 144, pp. 41-58.

Rob, S. et al. 2008. How to Avoid Colloquial (Informal) Writing. WikiHow.

Siddiqi, Anis. 2008. The Difference Between Formal and Informal Writing. EzineArticles.

Stamatatos, E., Michos, S., Fakotakis, N., and Kokkinakis, G. 1997. A User-Assisted Business Letter Generator Dealing with Text's Stylistic Variations. The Ninth International Conference on Tools with Artificial Intelligence, (TAI-97).

Woods, Geraldine. 2010. English Grammar for Dummies. 2nd Edition, page 147, Wiley Publishing, Inc. NJ, USA.

Zapata, Argenis A. 2008. Inglés IV (B-2008). Universidad de Los Andes, Facultad de Humanidades y Educación, Escuela de Idiomas Modernos.