Domain-specific variation of sentiment expressions: a methodology of analysis for academic writing

Stefania Degaetano-Ortlieb
Saarland University
s.degaetano@mx.uni-saarland.de

Elke Teich
Saarland University
e.teich@mx.uni-saarland.de

Ekaterina Lapshinova-Koltunski
Saarland University
e.lapshinova@mx.uni-saarland.de

Abstract

In this paper, we present work in progress towards a methodology for the analysis of domain-specific sentiment. In our case, we consider highly specialized scientific disciplines at the boundaries of computer science and selected other disciplines (e.g., computational linguistics, bioinformatics). Our approach is corpus-based and comprises the detection, extraction and annotation of features related to sentiment expressions, focusing on opinion targets.

1 Introduction

While many studies have been dedicated to the exploration of sentiment expressions, there is no comprehensive or uniform method of analysis. Studies vary not only in terms of their methodological (text-based, corpus-based, computational) but also in their theoretical approaches, see e.g., appraisal (Martin and White, 2005; Bednarek, 2006; Hood, 2010), stance (Biber and Finegan, 1989; Hyland, 2005), evaluation (Hunston and Thompson, 2001; Hunston, 2011) and sentiment analysis (Wilson, 2008; Somasundaran, 2010; Taboada et al., 2011). This multifaceted picture is also due to the phenomenon itself, which can be realized in various linguistic ways, especially when considering different domains.

We introduce a methodology for a semi-automatic corpus-based analysis of features related to sentiment expressions in academic writing. The methodology comprises: (1) feature detection by a manual annotation of features related to sentiment expressions in a small corpus of 100.000 tokens, (2) automatic feature extraction for comparative analyses of domain-specific variation in a corpus of 34 million tokens, and (3) automatic feature annotation to enrich the corpus.

2 Theoretical framework and data

Our methodology of analysis is based on the theoretical framework of Systemic Functional Linguistics (SFL) (Halliday, 2004). In SFL each element in a language is explained by reference to its function in the overall linguistic system and components of meaning are seen as functional components. There are three main functional components inherent to language, i.e. metafunctions: ideational (entities involved in a discourse), interpersonal (personal participation) and textual (structure of information). The core entities involved in sentiment expressions are writer and reader as well as target, which relates to any entity evaluated in a discourse. Sentiment expressions and the relation between a writer and a reader belong to the interpersonal metafunction.

With its register theory, SFL also accounts for domain-specific variation, assuming that meaning is realized in language by specific lexico-grammatical features, and different distributions of these features give rise to specific registers. As we are interested in domain-specific variation of sentiment expressions in registers emerged by register contact (e.g., bioinformatics emerged by contact between computer science and biology), we compare the distribution of interpersonal lexico-grammatical features within what we call *contact registers* (such as bioinformatics) and *seed registers* (such as computer science and bi-

core entities	examples of features	realization examples
writer	self-mention epistemic expressions	*I, we, our* *possible, may, suggest*
reader	reader pronouns directives	*you, your* *consider, see*
target	TARGET_BE_eval-expr TARGET_eval-V	*The claim is true.* *We see that A fails to be a BPP algorithm.*

Table 1: Interpersonal lexico-grammatical features

ology). These features can then be related back to the core entities inherent to sentiment expressions. In the case of academic writing, we have writer- and reader-oriented features, which relate to the writer's presence and the reader engagement in a discourse (Hyland, 2005; Biber et al., 1999), as well as target-oriented features, which relate to the entity evaluated (see Table 1 for examples). So far, studies of sentiment expressions in academic writing have mainly focused on writer-oriented and reader-oriented features (also known as stance and engagement features, see Degaetano and Teich (2011), Hyland (2005) and McGrath and Kuteeva (2012)). In this paper, we focus on target-oriented lexico-grammatical features in academic writing.

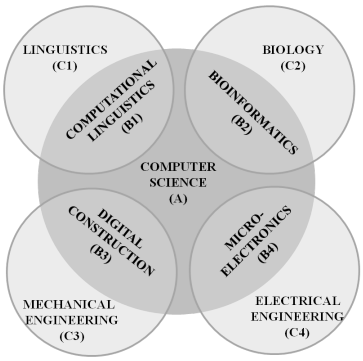

Figure 1: SciTex corpus

For our investigation we use the English Scientific Text corpus (SciTex; see Figure 1), specifically built to analyze register contact (Teich and Fankhauser, 2010; Degaetano-Ortlieb et al., 2012). SciTex contains a subcorpus for the contact registers (computational linguistics, bioinformatics, digital construction and microelectronics) and two subcorpora for the seed registers (one for computer science and one for linguistics, biology, mechanical engineering and electrical engineering). The corpus amounts to approx. 34 million tokens and is segmented into sentences, tokenized, lemmatized and part-of-speech tagged.

3 Methods

Our methodology for the analysis of sentiment expressions comprises a threefold semi-automatic process: detection, extraction, and annotation of lexico-grammatical features.

To detect features related to sentiment expressions, we look at a subcorpus of SciTex (approx. 100.000 tokens) and manually annotate interpersonal features related to writer, reader and target for each register. For this purpose, we use the UAM CorpusTool (O'Donnell, 2008), which allows to build an own annotation scheme and to adapt the scheme during annotation. Out of the annotation scheme, we generate a list of writer-, reader-, and target-oriented features.

The list of interpersonal features as well as insights gained from the manual annotation on different linguistic realizations of these features serve to create rules for the automatic extraction of features. For the extraction, we use the Corpus Query Processor (CQP), part of the IMS Corpus Workbench (CWB) (Evert, 2005; CWB, 2010), which allows feature extraction in terms of regular expressions over tokens and their linguistic annotations (e.g., part-of-speech). Especially for the extraction of target-oriented features, macros are built that cover several linguistic realizations of these features based on the insights of the manual annotation. As an example, Table 2 shows an extract of the macro for the *TARGET_BE_eval-expr* feature, which extracts realizations with and without a relative pronoun (compare the realization examples in Table 2). As not every adjective following the sequence *target+verb-BE* is evaluative, we use the lexical constraint *$eval-adj* (see

	query building blocks	comments	realization
1	MACRO TARGET_BE_eval-expr(0) (	begin macro and first query	
2	[pos="N.*"]	noun	*results*
3	[pos="MD"]?	optional modal verb	-
4	[lemma="be"]	verb BE	*are*
5	[pos="RB"]?	optional adverb	*quite*
6	[word=$eval-adj]	evaluative adjective	*good*
7	) | (	end first and begin second query	
8	[]	any token	*terminology*
9	[word=","]?	optional comma	-
10	[pos="WDT"]	relative pronoun	*that*
11	[pos="MD"]?	optional modal verb	*will*
12	[lemma="be"]	verb BE	*be*
13	[pos="RB"]?	optional adverb	
14	[word=$eval-adj] ...	evaluative adjective	*useful*

Table 2: Extract of the macro for the target-oriented feature *TARGET_BE_eval-expr*

lines 6 and 14 in Table 2), which restricts the extraction query to evaluative adjectives only, collected on the basis of the manual annotation. The automatic extraction is then performed on the full version of SciTex for comparative analyses of domain-specific variation.

Moreover, we want to automatically annotate the features back into the full version of SciTex to enrich the corpus with information at the interpersonal level. The annotation procedure is derived from the methods used in the YAC recursive chunker (Evert and Kermes, 2003) with CWB/CQP. The algorithm uses CWB perl-modules and perl-scripts to create additional annotation layers with CQP by writing back the query results of the extraction rules in form of new structural attributes into the corpus.

Based on the annotated features, we aim in the long-term at a corpus-based register analysis of interpersonal features. In the following section, we show selected analyses of target-oriented features.

4 Domain-specific variation - a sample analysis on target-specific tendencies

In order to analyze domain-specific variation of target-oriented features, we look at the differences in the targets evaluated across registers in SciTex as well as differences related to register contact by comparing contact registers with computer science or the other related seed register.

Considering the *TARGET_BE_eval-expr* feature, we observe domain-specific variation in terms of targets evaluated across registers. Table 3 shows a triple comparison of bioinformatics with computer science and biology for the five most frequent targets. The targets seem to be mostly domain-specific, especially for biology. Bioinformatics seems to adopt targets from both seed registers, e.g. *gene* and *algorithm*. The same holds for the other contact registers, e.g. *algorithm* is in the first twenty targets for computational linguistics and under the five most frequent targets for digital construction and microelectronics.

register (size)	target	F	per 1M
computer science (2,612,258)	*algorithm*	79	30.24
	problem	49	18.76
	result	45	17.23
	P	41	15.70
	lemma	31	11.87
bioinformatics (1,425,237)	*method*	50	35.08
	model	37	25.96
	gene	29	20.35
	algorithm	24	16.84
	approach	23	16.14
biology (2,161,297)	*gene*	41	18.97
	protein	38	17.58
	sequence	37	17.12
	region	30	13.88
	site	26	12.03

Table 3: Targets of the *TARGET_BE_eval-expr* feature

When we think of *algorithm* as a domain-specific target of computer science which is adopted by the contact registers, the question arises whether *algorithm* is evaluated in the same way, i.e. are the sentiment expressions also adopted from computer science or is *algorithm*

evaluated differently in the contact registers?

frames	comp. sci.		contact reg.	
	F	%	F	%
accuracy	2	2.82	2	2.56
being_necessary	2	2.82	1	1.28
correctness	6	**8.45**	2	2.56
desirability	21	**29.58**	9	11.54
difficulty	15	21.13	14	17.95
importance	2	2.82	8	**10.26**
likelihood	4	5.63	4	5.13
obviousness	2	2.82	0	0.00
sufficiency	2	2.82	2	2.56
suitability	4	5.63	4	5.13
usefulness	10	14.08	23	**29.49**
success	0	0.00	4	5.13
fame	0	0.00	2	2.56

Table 4: *TARGET_BE_eval-expr*: eval. frames with *algorithm*

To answer this question, we extract the sentiment expressions used to evaluate *algorithm* with CQP and categorize them semantically to have a basis of comparison. The semantic categorization is done manually according to FrameNet (Ruppenhofer et al., 2010), which provides a dataset of semantic frames of lexical units. The sentiment expressions are inserted in the online FrameNet interface which outputs the respective frames for each expression. Ultimately, we aim an automatic categorization. Table 4 shows semantic frames used in the *TARGET_BE_eval-expr* feature in computer science and the contact registers. The percentages show that computer science uses *algorithm* more frequently with the frames desirability (e.g., *perfect, good*) and correctness (by *correct*), the contact registers instead more frequently with importance (e.g., *key element*, *important*) and usefulness (e.g., *helpful*, *useful*).

To test whether these differences are significant, we calculate the fisher's exact test, a univariate method for small raw frequencies, with the R environment (R Development Core Team, 2010) for the relevant frames (desirability, correctness, importance and usefulness). The p-value for this comparison is 0.00119, showing that computer science and the contact registers differ significantly in their use of these frames with *algorithm* in the *TARGET_BE_eval-expr* feature.

However, while the expression of correctness may be an evaluative act in some domains,to call an *algorithm* 'correct' in computer science is rather a factual attribution. Thus, the concepts used to evaluate may vary according to register. We have to investigate further the semantic diversification related to scientific registers as well as the appropriateness of FrameNet frames for academic concepts.

Nevertheless, in summary we can say that for the *TARGET_BE_eval-expr* feature the contact registers adopt targets from the seed register computer science, but in the case of *algorithm* evaluate it differently.

5 Summary and envoi

In this paper, we have introduced a methodology to analyze sentiment expressions in academic writing. Our focus of analysis has been on specific lexico-grammatical features used to evaluate targets. As our overarching goal is to analyze registers emerged by register contact, we have (1) analyzed whether opinion targets are adopted from the seed register by the contact register, and (2) whether these targets are evaluated in the same way by both seed and contact registers. The analysis shows that there are domain-specific targets adopted by the contact registers. However, the evaluation of the targets can differ, as in the case of *algorithm*. Thus, we were able to detect domain-specific variation in terms of sentiment expressions, i.e. for interpersonal lexico-grammatical features.

In terms of methods, we have applied a corpus-based approach that allows to detect, extract and annotate features related to sentiment expressions semi-automatically in academic writing.

In the future, we will widen the range of target-oriented features moving towards a more comprehensive picture of domain-specific variation in terms of targets. We also have to take into account the semantic diversification related to different registers and investigate further the appropriateness of FrameNet categories to describe concepts used within scientific writing. As our model of analysis accounts also for writer- and reader-oriented features, we investigate these features as well. Doing so, we aim at analyzing domain-specific variation in terms of the strength of the writer's presence in specific domains and the reader's engagement across registers.

References

Monica Bednarek. 2006. *Evaluation in Media Discourse: Analysis of a Newspaper Corpus*. Continuum.

Douglas Biber and Edward Finegan. 1989. Drift and the evolution of English style: a history of three genres. *Language*, 65(3):487–517.

Douglas Biber, Stig Johansson, and Geoffrey Leech. 1999. *Longman Grammar of Spoken and Written English*. Longman, Harlow.

2010. The IMS Open Corpus Workbench. http://www.cwb.sourceforge.net.

Stefania Degaetano and Elke Teich. 2011. The lexicogrammar of stance: an exploratory analysis of scientific texts. In Stefanie Dipper and Heike Zinsmeister, editors, *Bochumer Linguistische Arbeitsberichte 3 - Beyond Semantics: Corpus-based Investigations of Pragmatic and Discourse Phenomena*, Bochum.

Stefania Degaetano-Ortlieb, Kermes Hannah, Ekaterina Lapshinova-Koltunski, and Teich Elke. 2012. SciTex A Diachronic Corpus for Analyzing the Development of Scientific Registers. In Paul Bennett, Martin Durrell, Silke Scheible, and Richard J. Whitt, editors, *New Methods in Historical Corpus Linguistics*, volume Corpus Linguistics and Interdisciplinary Perspectives on Language - CLIP, Vol. 3. Narr.

Stefan Evert and Hannah Kermes. 2003. Annotation, storage, and retrieval of mildly recursive structures. In *Annotation, storage, and retrieval of mildly recursive structures*, Lancaster, UK.

Stefan Evert, 2005. *The CQP Query Language Tutorial*. IMS Stuttgart. CWB version 2.2.b90.

M.A.K. Halliday. 2004. *An Introduction to Functional Grammar*. Arnold, London.

Susan Hood. 2010. *Appraising Research: Evaluation in Academic Writing*. Palgrave Macmillan.

Susan Hunston and Geoff Thompson. 2001. *Evaluation in Text: Authorial stance and the construction of discourse*. Oxford University Press, Oxford.

Susan Hunston. 2011. *Corpus approaches to evaluation: phraseology and evaluative language*. Taylor & Francis.

Ken Hyland. 2005. Stance and engagement: a model of interaction in academic discourse. *Discourse Studies*, 7(2):173–192.

Jim R. Martin and Peter R.R. White. 2005. *The Language of Evaluation, Appraisal in English*. Palgrave Macmillan, London & New York.

Lisa McGrath and Maria Kuteeva. 2012. Stance and engagement in pure mathematics research articles: Linking discourse features to disciplinary practices. *English for Specific Purposes*, 31:161–173.

Michael O'Donnell. 2008. The UAM CorpusTool: Software for corpus annotation and exploration. In *Proceedings of the XXVI Congreso de AESLA*, Almeria, Spain, 3-5 April.

R Development Core Team, 2010. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Josef Ruppenhofer, Michael Ellsworth, Miriam R. L. Petruck, Christopher R. Johnson, and Jan Scheffczyk. 2010. Framenet II: Extended theory and practice. Technical report, ICSI.

Swapna Somasundaran. 2010. *Discourse-level relations for opinion analysis*. Ph.D. thesis, University of Pittsburgh.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307.

Elke Teich and Peter Fankhauser. 2010. Exploring a corpus of scientific texts using data mining. In S. Gries, S. Wulff, and M. Davies, editors, *Corpus-linguistic applications: Current studies, new directions*, pages 233–247. Rodopi, Amsterdam and New York.

Theresa Ann Wilson. 2008. *Fine-grained Subjectivity and Sentiment Analysis: Recognizing the Intensity, Polarity, and Attitudes of Private States*. Ph.D. thesis, University of Pittsburgh.