Analysing the Integration of Semantic Web Features for Document Planning across Genres

Marta Vicente
Department of Software
and Computing Systems
University of Alicante
Apdo. de Correos 99
E-03080, Alicante, Spain
mvicente@dlsi.ua.es

Elena Lloret
Department of Software
and Computing Systems
University of Alicante
Apdo. de Correos 99
E-03080, Alicante, Spain
elloret@dlsi.ua.es

Abstract

Language is usually studied and analysed from different disciplines generally on the premise that it constitutes a form of communication which pursues a specific objective. The discourse, in that sense, can be understood as a text which is constructed to express such objective. When a discourse is created, its production is related to some textual genre, usually connected with some pragmatic features, like the intention of the writer or the audience to whom is addressed, both conditioning the use of language. But genres can be considered as well as compounds of different pieces of text with a certain degree of order, each one seeking for more concrete objectives. This paper presents a proposal to learn such features as a way to generate richer document plans, applying clustering techniques over annotated documents.

1 Motivation and Research Context

The current research is carried out from a conception of Natural Language Generation (*NLG*) for which the creation of a text requires an intermediate output called a document plan. It is by the macroplanning stage that the system provides this plan of selected and ordered content. At present, our work is focused on how to elaborate that plan in order to meet some requisites regarding flexibility of the system: it should be able to produce different outcomes conditioned by the communicative goal, the audience,... the context, on the whole. Henceforth, the main aim of our current research is to enrich the pragmatic facet of the *NLG* process. The expected outcome is a scheme or ordering of the ideas that should be realised in a set of cohesive and coherent sentences and paragraphs.

According to some theories of the discourse (Bakhtin, 2010; Halliday et al., 2014), genres can be understood as social constructions that settle a connection between the discourse and the situation in which it is produced, reflected both in its structure and its content. According to Swavels (1990):

> "A genre comprises a class of communicative events, the members of which share some set of communicative purposes. These purposes are recognised by the expert members of the parent discourse community, and thereby constitute the rationale for the genre. This rationale shapes the schematic structure of the discourse and influences and constraints choice of content and style."

Besides, genres become interesting because they are related to communicative purposes in different manners, from a global viewpoint to fine-grained levels. As an example, we can think on the case of a person who is looking for recommendation in review pages. Recommending would be the main, global purpose of the text he consults when it was created. But it is possible that the writer also wanted to explain the motivation of the journey - narrative, personal experience - or to describe the facilities in order to complete his review. Narration, description, recommendation,... they represent low-level functions of the text related to the intention of the writer and, in some cases, they can be identified as different sets of sentences. This lead us to the possibility of learning the structure of the text and its features, which differs from one genre to another. In reviews, the presence and order of the parts is not strict.

Maybe one traveller does not share his personal story, but also he describes the room and recom-

mends the brand, while another one first evaluates and then describes. An example to illustrate this can be found in table 1. Conversely, it would make no sense to write a scientific article that reports the results before explaining the methodology or not explaining it at all, for example.

Review 1
Personal Experience: *On our last trip to Hawaii my husband and I...* *As an added bonus, we were given...* *We decided to take advantage of...*
Description: *The lobby is adorned with lush gardens...* *Alongside the gardens are tropical birds...* *The rooms are spacious.*
Recommendation: *If you are ever fortunate enough to visit* *the beautiful island of Kauai, try to stay* *at the H Regency, you won't be disappointed.*
Review 2
Description: *The W New York is on Lexington right...* *The rooms are just as small as before...* *The lobby of the hotel is also...*
Personal Experience: *Being a corporate lawyer I travel...* *The first time I was in a small room...* *The second time I could not believe...*
Description: *Although the room size is awful,* *the hotel does have some nice touches.* *Another benefit of the hotel is that...*

Table 1: Review ordering from a functional approach. Just with the first words of the sentences some characteristic features can be appreciated (Verb tenses, person-thirdfirst-, ...)

Therefore, our hypothesis is that it is possible to characterise subparts of a discourse (related to a genre) according to their functionality and, at the same time, learn about its (flexible) ordering. Due to the lack of annotated corpora with discourse information about that communicative purposes, we propose to work with unsupervised techniques to achieve that goal. We expect to obtain the necessary knowledge to produce appropriate document plans. Taking into account several genres that normally exhibit a pre-defined or known-in-advance structure, such as the case of news, Wikipedia pages, or scientific article, we would be able to validate our suggested approach in other textual genres that lacks such well-defined structure a priori.

Our methodology relies on pattern detection techniques. Until now, we have tried clustering that does not require previous knowledge of the number of clusters. Over an annotated corpus we apply an Expectation-Maximisation (*EM*) algorithm, having included within the features linguistic information related to its placement.

The remainder of this paper is organised as follows. Section 2 summarises the related work concerning text classification efforts and genre studies related to communication objectives. Section 3 describes the kind of linguistic lexical features that we have been using in our experiments until now. After that, section 4 describes some resources coming from the Semantic Web environment that could complement and enrich those features. Finally, section 5 describes the experiments already performed and outlines future research opportunities.

2 Related Work

Back in 1997, Hearst tried to detect the structure of text using patterns of lexical co-occurrence to identify paragraphs related to the same topic (Hearst, 1997). In this case, term repetition proved to be enough to detect subtopics in explanatory texts, but did not include consideration about other traits of the discourse (e.g. syntactic constructions, verb tenses, number of adjectives in each region) neither recovering more meaning further than topic identification, as could be the purpose intended on the paragraph(s). Besides, the author remarked that the results had proved highly valuable when applied to explanatory text, but they would be less significant for other text types.

From another point of view, Bachand (Bachand et al., 2014) develops a research focused on the relations between text-type, discourse structures and rhetorical relations. Again, the experiments conducted are implemented on a single type of feature, this time rhetorical relations and markers. The good results obtained by the author indicate that our approach, which is grounded in similar intuitions, can reach comparable developments that we expect will enrich our capacity for generating accurate document plans.

Regarding reviews, most of the work developed refers to sentiment analysis or polarity classifica-

tion (Cambria et al., 2013). A few research works have been focused on the structure related to textual genres, relying on the Systemic Functional Theory (Taboada, 2011). The relations of different parts of the text with several purposes are revealed, focusing their analysis on the domain of movie reviews, and showing at the same time the variability of the ordering in such type of documents.

Finally, a special mention must be done to the Systemic Functional Theory (Halliday et al., 2014). It provides a notion of genre that connects situation types with semantic/lexico-grammatic patterns from a conception of language highly related to its socio-semiotic origin. A textual typology is depicted on this terms, connected as well with the context of the discourse and the semantic choices to organise it (Matthiessen, 2014). On the other hand, and as a more precise example, the typology of processes that Halliday and Mathiessen describe, directly influences the classification accomplished by *ADESSE*, one of the resources applied in our experiments over Spanish reviews, explained in the next section.

3 Analysis of the Features

Having pointed out the expanse of the related work, our approach wants to overcome its limitations. On the one hand, in the sense of being suitable for any genre, not a particular one. On the other hand, focusing on several types of features at the same time, in order to propose a more comprehensive description of the parts of a discourse.

With regard to accomplish such a project, the selection and design of the proper features becomes a challenging task itself, strongly related to the aim of the investigation. Specifically, we try to detect the features that may reveal links with the functionality or purpose of the paragraph that includes them. We have begun annotating several aspects by means of linguistic tools and resources: *Freeling* (Padró and Stanilovsky, 2012) for PoS annotation and Entity Recognition and *ADESSE* (García-Miguel et al., 2010) as a source of verb senses from a semantic perspective.

4 Semantic Web to enrich the Data Set

We believe that, in order to become more meaningful, the quality of features could be improved by means of some resources rooted in Web Semantic technologies. There is some research related to genres that can be useful in our project. In the realm of reviews, opinion and sentiment annotation, we can take advantage for example of *MARL Ontology Specification*[1], a data schema that has been used in the *EuroSentiment Project* (Buitelaar et al., 2013) or directly related to reviews from a Sentiment Analysis perspective (Santosh and Vardhan, 2015). Other genres have been targeted for similar developments. With regard to news genre, in order to obtain more significant annotation of the documents, *BBC* provides a set of ontologies related to their contents. *DBPedia* has been already proved useful for Wikipedia articles researchers. *Drammar* (Lombardo and Damiano, 2012) and *OntoMedia* (Jewell et al., 2005) are ontology-based models for annotating features of media and cultural narratives. All of them represent resources that may lead to different results in our clustering task and analysis.

ADESSE verb senses
Mental, material, relational, verbal, existential and modulation
FREELING features
PoS tagging: noun, adjective, pronoun, verb (tense, aspect, ...), etc.

Table 2: Features annotated over the corpus of reviews.

5 On-going Work

Until now, some experiments have been performed over a corpus of Spanish reviews extracted from Tripadvisor. The reviews were segmented into sentences, and some figures regarding semantic and morphological features were computed after dividing each document in regions (sets of sentences), increasing their number from one block up to four blocks of sentences. Table 3 shows some statistics of the corpus employed.

Number of reviews	1400
Sentences	12,467
Words labelled	around 200,000

Table 3: Corpus statistics.

In order to strengthen the results, corpora of other genres with different degree of flexibility in their structure are being analysed: tales, news and Wikipedia articles are to be compared with the for-

[1] http://www.gsi.dit.upm.es/ontologies/marl

mer outcomes. The length of the blocks is the result of a proportional division of the length of the document for now. As the research advances, new experiments will be developed to determine a more accurate size for the pseudo-paragraphs. With the ideas introduced in the section 4, our next step and proposal, includes improving the significance of the features with which the clustering algorithms have to work, trying to reveal an inner structure of the text related to its genre and purposes. The better our features are, the more precise the descriptions we can do of the discourse areas.

Acknowledgments

This research work has been supported by the Generalitat Valenciana by the grant ACIF/2016/501.It has also funded by the University of Alicante, Spanish Government and the European Commission through the projects, "Explotación y tratamiento de la información disponible en Internet para la anotación y generación de textos adaptados al usuario" (GRE13-15) and "DIIM2.0: Desarrollo de técnicas Inteligentes e Interactivas de Minería y generación de información sobre la web 2.0" (PROMETEOII/2014/001), TIN2015-65100-R, TIN2015-65136-C2-2-R, and SAM (FP7-611312), respectively.

References

Félix-Hervé Bachand, Elnaz Davoodi, and Leila Kosseim. 2014. An investigation on the influence of genres and textual organisation on the use of discourse relations. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 454–468. Springer.

Mikhail Mikhaĭlovich Bakhtin. 2010. *Speech genres and other late essays*. University of Texas Press.

Paul Buitelaar, Mihael Arcan, Carlos Angel Iglesias Fernandez, Juan Fernando Sánchez Rada, and Carlo Strapparava. 2013. Linguistic linked data for sentiment analysis.

Erik Cambria, Bjorn Schuller, Yunqing Xia, and Catherine Havasi. 2013. New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems*, 28(2):15–21.

José M. García-Miguel, Gael Vaamonde, and Fita González Domínguez. 2010. Adesse, a database with syntactic and semantic annotation of a corpus of spanish. In *LREC*.

MAK Halliday, Christian MIM Matthiessen, Michael Halliday, and Christian Matthiessen. 2014. *An introduction to functional grammar*. Routledge.

Marti A. Hearst. 1997. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational linguistics*, 23(1):33–64.

Michael O. Jewell, K. Faith Lawrence, Mischa M. Tuffield, Adam Prugel-Bennett, David E. Millard, Mark S. Nixon, Nigel R. Shadbolt, et al. 2005. Ontomedia: An ontology for the representation of heterogeneous media. In *In Proceeding of SIGIR workshop on Mutlimedia Information Retrieval*. ACM SIGIR.

Vincenzo Lombardo and Rossana Damiano. 2012. Semantic annotation of narrative media objects. *Multimedia Tools and Applications*, 59(2):407–439.

Christian MIM Matthiessen. 2014. Registerial cartography: context-based mapping of text types and their rhetorical-relational organization.

Lluís Padró and Evgeny Stanilovsky. 2012. FreeLing 3.0: Towards Wider Multilinguality. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. European Language Resources Association (ELRA).

D. Teja Santosh and B. Vishnu Vardhan. 2015. Feature and sentiment based linked instance rdf data towards ontology based review categorization. In *Proceedings of the World Congress on Engineering*, volume 1.

John Swales. 1990. *Genre analysis: English in academic and research settings*. Cambridge University Press.

Maite Taboada. 2011. Stages in an online review genre. *Text & Talk-An Interdisciplinary Journal of Language, Discourse & Communication Studies*, 31(2):247–269.