The Content Types Dataset: a New Resource to Explore Semantic and Functional Characteristics of Texts
Sprugnoli, Rachele; Caselli, Tomasso; Tonelli, Sara; Moretti, Giovanni

Published in:

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2017

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.
The Content Types Dataset: a New Resource to Explore Semantic and Functional Characteristics of Texts

Rachele Sprugnoli\textsuperscript{1,3}, Tommaso Caselli\textsuperscript{2}, Sara Tonelli\textsuperscript{1} and Giovanni Moretti\textsuperscript{1}

\textsuperscript{1}DH-FBK, Via Sommarive 18, Povo, Trento
\textsuperscript{2}CLTL, Vrije Universiteit Amsterdam, De Boelelaan, 1105, Amsterdam
\textsuperscript{3}University of Trento, Via Sommarive 9, Povo, Trento
\{sprugnoli, satonelli, moretti\}@fbk.eu
\{t.caselli\}@vu.nl

Abstract

This paper presents a new resource, called Content Types Dataset, to promote the analysis of texts as a composition of units with specific semantic and functional roles. By developing this dataset, we also introduce a new NLP task for the automatic classification of Content Types. The annotation scheme and the dataset are described together with two sets of classification experiments.

1 Introduction

This paper introduces a new resource and task for NLP, namely the classification of Content Types. The notion of Content Types differs from standard discourse relations, either based on rhetorical structures or lexically-grounded approaches. Content Types provide cues to access the structure of a document’s types of functional content. They contribute to the overall message or purpose of a text and make explicit the functional role of a discourse segment with respect to its content, i.e. meaning. Their identification may improve the performance of more complex NLP tasks by targeting the portions of the documents that are more relevant. For example, when building a storyline it may be useful to focus on the narrative segments of a text (Vossen et al., 2015), while for sentiment analysis the identification of evaluative clauses may be beneficial (Liu, 2015).

Our contribution is threefold: i) we make available annotation guidelines with high reliability in terms of inter-annotator agreement and applicable to texts of different genres and period of publication; ii) we release the first version of a new dataset (whose annotation is still in progress) that takes into consideration both contemporary and historical texts, paving the way to a new NLP task, i.e. Content Type Classification; and iii) we present initial promising results for the automatic classification of Content Types by using the first version of the dataset. All data are made available on-line\textsuperscript{1}.

The remainder of the paper is structured as follows: Section 2 illustrates the annotation scheme, the composition of the dataset, and report the inter-annotator agreement. Section 3 presents two sets of experiments to automatically classify Content Types. Related work is discussed in Section 4. Finally, conclusion and future work are reported in Section 5.

2 Dataset Construction

Content Types (henceforth CTs) are text passages with specific semantic and functional characteristics. Their definition is based on linguistic features, and the annotation is performed at clause level. Clauses are considered as textual constituent units (Polanyi, 1988), and defined as groups of words related to each other, containing a finite or non-finite verb, while the subject may be implicit or shared with other clauses. This granularity level of the mark-up was chosen to provide a fine-grained annotation of CTs that can characterize different portions of the same sentence. Example (1) is made of two clauses (divided by “//”): the first narrates what the author is doing, the second describes the place where she is.

\begin{quote}
(1) I am writing on a fine terrace overlooking the sea,// where stone benches and tables are conveniently arranged for our use.
\end{quote}

We identify seven classes of CTs, five of which are based from Werlich’s typology, while the last two (OTHER and NONE) were introduced in our

\textsuperscript{1}\url{https://github.com/dhfbk/content-types}
To test the comprehensiveness of this scheme, we annotate English texts from two different genres and periods of publication: namely, contemporary news and travel reports published between the end of the XIX Century and the beginning of the XX Century. While the former are taken from already available datasets, i.e., TempEval-3, Penn Discourse Treebank, and MASC (UzZaman et al., 2013; Prasad et al., 2008; Ide et al., 2010), the latter constitute a novel set of texts extracted from the Gutenberg project\(^2\). The corpus is released under the name of Content Types Dataset version 1.0 (CTD_v1). The resource is still being extended with new annotated texts, but in the remainder of the paper we will refer to this first version.

The annotation was conducted by two expert linguists following a multi-step process and using the web-based tool CAT (Bartalesi Lenzi et al., 2012). In the first phase, annotators were allowed to discuss disagreements based on a trial corpus suggesting revisions to improve the guidelines. In the second phase, inter-annotator agreement was calculated on a subset of the CTD_v1 (a total of 5,328 tokens and 526 clauses, with 2,500 tokens and about 250 clauses per genre). Table 1 reports the Cohen’s kappa on the number of tokens for both text genres. With the exception of the INSTRUCTIVE CT, all the classes have high scores, exceeding 0.8, usually set as a threshold that guarantees good annotation quality (Artstein and Poesio, 2008). In the final phase, the whole dataset was annotated using the latest version of the guidelines which includes detailed descriptions of the classes, examples for both genres, and priority rules discriminating when more than one CT class may apply to clauses. Table 2 illustrates the composition of CTD_v1. The two genres of texts show, for almost all the CT classes, a statistically significant difference (at \(p<0.01\) and calculated with the \(z\) test) in their distribution.

### Table 2: Composition of CTD_v1

<table>
<thead>
<tr>
<th>Class</th>
<th>News</th>
<th>Travel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evaluative</td>
<td>0.82</td>
<td>0.90</td>
</tr>
<tr>
<td>Descriptive</td>
<td>0.84</td>
<td>0.86</td>
</tr>
<tr>
<td>Expository</td>
<td>-</td>
<td>0.93</td>
</tr>
<tr>
<td>Instructive</td>
<td>-</td>
<td>0.65</td>
</tr>
<tr>
<td>Narrative</td>
<td>0.86</td>
<td>0.88</td>
</tr>
<tr>
<td>None</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>Other</td>
<td>-</td>
<td>0.92</td>
</tr>
</tbody>
</table>

3 Experiments

In this section we present initial experiments for the automatic classification of clauses in CTS. Attribute classification was not targeted at this stage. We conducted two sets of experiments to test different modeling assumptions. In all experiments we use gold clause boundaries.

\(^2\)http://www.gutenberg.org/
Table 2: Statistics of CTD_v1: an asterisk indicates whether the content type has a statistically significant difference in the distribution over the two genres.

<table>
<thead>
<tr>
<th>Content Type</th>
<th>News</th>
<th>Travel Reports</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Texts</td>
<td>Tokens</td>
<td>Clauses</td>
</tr>
<tr>
<td>Evaluative*</td>
<td>428 (14.09%)</td>
<td>618 (19.59%)</td>
<td>1,046 (16.88%)</td>
</tr>
<tr>
<td>Descriptive*</td>
<td>198 (6.52%)</td>
<td>480 (15.19%)</td>
<td>678 (10.94%)</td>
</tr>
<tr>
<td>Expository</td>
<td>58 (1.91%)</td>
<td>81 (2.56%)</td>
<td>139 (2.24%)</td>
</tr>
<tr>
<td>Instructive</td>
<td>5 (0.16%)</td>
<td>4 (0.13%)</td>
<td>9 (0.15%)</td>
</tr>
<tr>
<td>Narrative*</td>
<td>2,318 (76.30%)</td>
<td>1,738 (55.03%)</td>
<td>4,056 (65.46%)</td>
</tr>
<tr>
<td>None*</td>
<td>15 (0.49%)</td>
<td>38 (1.20%)</td>
<td>53 (0.86%)</td>
</tr>
<tr>
<td>Other*</td>
<td>16 (0.53%)</td>
<td>199 (6.30%)</td>
<td>215 (3.47%)</td>
</tr>
</tbody>
</table>

Table 3: Features of the clause components.

3.1 Feature Sets

We experiment two different types of features: the first relies on distributional information extracted through sentence embeddings (Le and Mikolov, 2014), while the second is linguistically motivated and focuses on syntactic and semantic properties of the main components of the clause, i.e. the noun phrase(s) and the verb phrase. For the first type, we extracted embeddings for each clause using the doc2vec (Le and Mikolov, 2014) implementation in gensim, with vector size = 50 and window = 5. For the second feature type, all documents were pre-processed at clause level with Stanford CoreNLP (Manning et al., 2014), performing tokenization, lemmatization, POS tagging, Named Entity recognition. The extraction of basic syntactic and semantic properties of the clause components has been performed with a syntactic-semantic features toolkit (Friedrich and Pinkal, 2015). This has allowed us to identify four blocks of features for: (i) the noun phrase in subject position (i.e. nsubj and nsubjpass), (ii) the noun phrase in direct object position (i.e. dobj and agent), (iii) the noun phrase in any other syntactic relation, and (iv) the clause verb. Details for noun phrase and verb phrase components are reported in Table 3.

We extended the basic features with prior sentiment polarity scores for nouns, verbs, adjectives, and adverbs in the clause via SentiWordNet (Baccianella et al., 2010). For each target POS, polarity scores are aggregated per lemma and averaged by the number of senses, thus providing a lemma-based prior polarity. Finally, the lemma-based polarity scores are normalized by the clause length and scaled between 0 and 1. Finally, we introduced a binary feature to mark the presence/absence of a temporal expression in a clause. These two additional blocks of features have been selected following the definition of the CTs in the annotation guidelines. In particular, the presence of temporal expressions in a clause can facilitate the distinction between the NARRATIVE and the DESCRIPTIVE classes, while the polarity features should facilitate the identification of the EVALUATIVE class.
3.2 Classification Experiments

We developed our models by dividing the annotated data in training (80%) and test sections (20%), balancing the distribution in each section across the two genres. The overall amount of clauses in the training and test data is slightly lower than the one of the manually annotated clauses\(^3\): indeed, we excluded some clauses because the pre-processing tools were not able to extract any relevant features from them. This is mainly due to a failure of the syntactic-semantic toolkit to process some gold clauses.

To better evaluate the performance of our models, we developed a baseline system by assigning the most frequent CT per text genre on the basis of the frequencies in the training data. Evaluation has been computed by means of Precision, Recall, and F1-score as implemented in scikit-learn (Pedregosa et al., 2011).

### Content-based Classification

In this set of experiments we aimed at verifying the fitness of our features by assuming that CTs are independent of each other and determined only by their meaning. We developed four classifiers, by varying the combination of features, using two different learners, namely Support Vector Machines (SVM) (Cortes and Vapnik, 1995) and Conditional Random Fields (CRFs) (Lafferty et al., 2001):

- **clause model** has only basic clause features plus the polarity scores and the presence/absence of temporal expressions.
- **clause+doc2vec model** has the clause model feature set extended with the doc2vec clause embeddings.

The SVM models have been implemented using LIBSVM (Chang and Lin, 2011) with Linear Kernel. The CRF models have been implemented with CRF++ toolkit \(^4\) with default parameters.

### Content and Functional Structure Classification

This set of experiments assumes an alternative modeling strategy by viewing each sentence as a sequence of CTs, each associated with a clause. For this second set of experiments we implemented two linear CRF classifiers by extending the previously described models with a context window of [+/−2] for all features.

Results are illustrated in Table 4. The content-based classification experiments show that CTs are subject to the functional structure of the sentence and, more generally, of the document. Only the CRF classifiers, i.e., sequence labeling models, can beat the baseline, providing balanced results for Precision and Recall, and improving the F1 score by 0.11 (CRF-clauseC) and 0.10 points (CRF-clause+doc2vecC). The SVM models, on the contrary, fail to beat the baseline. This could be due to the imbalanced distribution of CTs, and also to the fact that content features alone are not enough to discriminate the different CTs. The contribution of the doc2vec features is, however, limited: they help increasing the Recall values (+0.03 points) but have a little effect on the Precision (+0.01 point) when considering the CRF models. On the contrary, they do not provide any improvements with the SVM models.

As for the content and functional structure classification models, the results indicate that context features positively contribute to the improvement of the classification task (the CRF-clauseCF with context features outperforms its direct counterpart, CRF-clauseC, in the content-based classification setting). It is interesting to notice a redundancy between the doc2vec features and the context window. In this case, the CRF-clause+doc2vecCF has the lowest results for Precision and F1, and a slight increase in Recall (0.68 vs. 0.67).

4 Related Work

The classification of text passages has been studied in previous works considering different textual units (e.g., clauses, sentences, and paragraphs) or language patterns (Kaufer et al., 2004). Several annotation schemes, often based on genre-specific taxonomies, have been proposed. This is the case, for example, of the detection of the main components in scholarly publications (Teufel et al., 2009; Liakata et al., 2012; De Waard and Maat, 2012; Burns et al., 2016) or the annotation of content zones, i.e., functional constituents of texts (Bieler et al., 2007; Stede and Kuhn, 2009; Baiamonte et al., 2016). On the contrary, the notion of Content Types that we have adopted applies across genres. CTs are based on linguistic theories on discourse/rhetorical strategies but differ from discourse relations. Over the years, different typologies have been proposed (Werlich, 1976; Biber,
Table 4: Results of the classification experiments.

<table>
<thead>
<tr>
<th>Model</th>
<th>P</th>
<th>R</th>
<th>F1</th>
<th>Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (NARRATIVE)</td>
<td>0.42</td>
<td>0.65</td>
<td>0.51</td>
<td>0.65</td>
</tr>
<tr>
<td>SVM-clause</td>
<td>0.42</td>
<td>0.65</td>
<td>0.51</td>
<td>0.65</td>
</tr>
<tr>
<td>SVM-clause+doc2vec</td>
<td>0.42</td>
<td>0.65</td>
<td>0.51</td>
<td>0.65</td>
</tr>
<tr>
<td>CRF-clauseC</td>
<td>0.61</td>
<td>0.65</td>
<td>0.62</td>
<td>0.66</td>
</tr>
<tr>
<td>CRF-clause+doc2vecC</td>
<td>0.62</td>
<td>0.68</td>
<td>0.61</td>
<td>0.67</td>
</tr>
</tbody>
</table>

The theory of Discourse Modes (DMs) (Smith, 2003) is instead followed by Mavridou et al. (2015) that apply it to a paragraph-based pilot annotation of a variety of documents such as novels, news and European Parliament proceedings. Annotators intuitively labeled DMs relying on a very short manual: as a consequence, no formal guidelines were made available and only a moderate agreement was achieved. Moreover, the final dataset is not publicly available and the recognition of DMs has not been automated yet. Our approach is different: we rely on Werlich’s typology, we provide complete annotation guidelines, we make available the annotated dataset, and we experiment automatic classification of CTs.

5 Conclusion and Future Work

In this work, we presented a novel resource annotated with CTs and a set of experiments aimed at automatically classifying clauses based on content and on their functional structure. Although this work is still in progress, the proposed annotation scheme proved sound and the developed corpus can already provide insights into the functional role of discourse segments with respect to the clause meaning.

In addition to SVM and CRFs, we experimented with artificial neural networks (ANN) using the Keras\(^5\) framework running on the TensorFlow implementation (Abadi et al., 2015). We tested different configurations but results are not higher than those obtained with CRFs. We will investigate the reasons and try other models. Similarly, we will investigate whether SVM kernels other than the linear one can do better.

In the future, we will continue the annotation of the dataset, by introducing documents from other text genres (e.g. travel guides, news editorials, school textbooks) so as to re-balance the distributions of the CTs in the dataset. Furthermore, we plan to study whether information on content types can contribute to other NLP tasks. For example, we believe that identifying NARRATIVE and EVALUATIVE CTs may contribute to discriminating between clauses useful to build a storyline or a timeline of events (the former) and clauses bearing sentiment information (the latter).

6 Acknowledgement

Part of this work has been conducted during a visiting period of one of the author at the Computational Lexicology & Terminology Lab (CLTL) at the Vrije Universiteit Amsterdam. One of the author wants to thanks the NWO Spinoza Prize project Understanding Language by Machines (sub-track 3) for supporting this work.

References


\(^5\)https://github.com/fchollet/keras


