Using discursive information to disentangle French language chat

Matthieu Riou, Soufian Salim, Nicolas Hernandez

Using discursive information to disentangle French language chat

Matthieu Riou **Soufian Salim** **Nicolas Hernandez**
LINA UMR 6241 Laboratory
University of Nantes (France)
firstname.lastname@univ-nantes.fr

Abstract

In internet chatrooms, multiple conversations may occur simultaneously. The task of identifying to which conversation each message belongs is called disentanglement. In this paper, we first try to adapt the publicly available system of Elsner and Charniak (2010) to a French corpus extracted from the Ubuntu platform. Then, we experiment with the discursive annotation of utterances. We find that disentanglement performances can vary significantly depending on corpus characteristics. We also find that using discursive information, in the form of functional and rhetoric relations between messages, is valuable for this task.

1 Introduction

Interest in live chats has grown as they gained popularity as a channel for computer-mediated communication. While many chat services are designed to only allow dialogue in between two participants, a number of them let multiple participants join in and send messages into a common text stream. It is therefore frequent for multi-party chats to feature several simultaneous conversations. The task of identifying these conversations and the messages that belong to them is called disentanglement. It is a required preprocessing step for many higher-level analysis systems, such as those that rely on contextual knowledge about utterances. For example, dialogue act classifiers typically depend on information about previous utterances in the conversation (Kim et al., 2012). Moreover, any sequential analysis system, such as one based on CRFs, would in fact require chat disentanglement.

In this paper, we first consider the existing disentanglement system proposed by Elsner and Charniak (2010), which is based on lexical analysis, and attempt to adapt it to French language chats from the Ubuntu chatrooms. Then, we experiment with discursive information by annotating relations between messages, and try to see if adding such information to the feature sets improves upon the existing system.

2 Related Work

Elsner and Charniak are predominant in the literature. Most notably, they worked on the construction of an annotated corpus used as a reference for many works, *"Are you talking to me?"* (Elsner and Charniak, 2008). It was used by Wang and Oard (2009), who proposed a method for conversation reconstitution based on message context. Their results are state-of-the-art for Elsner and Charniak's corpus. Mayfield et al. (2012) proposed a learning model for the detection of information-sharing acts at the sentence level, then the aggregation of these sentences into sequences and finally the clustering of resulting sequences into conversations.

This paper is primarily based on the publicly available system presented by Elsner and Charniak (2010)[1]. They adopt a two-step approach to chat disentanglement. The first step is to determine for each pair of messages whether they belong to the same conversation or not. In order to do that, they start by identifying candidate message pairs. These pairs are formed based on whether the two messages were sent in an interval of 129 seconds or less[2]. The intuition behind this heuristic is that the more distant in time two messages are, the less likely it is that they belong to the same conversation. Then, they use a maximum-entropy classifier to determine whether they actually do. The second step is to partition these messages into several clusters to obtain the automatically annotated cor-

[1] http://www.ling.ohio-state.edu/~melsner/#software

[2] This particular value was chosen because it is the threshold after which the classifier no longer outperforms the majority baseline.

pus. They accomplish this by using a greedy voting algorithm.

Their system is based on lexical similarity between messages as well as chat-specific and discourse features. However, said discourse features remain simple and could easily be improved upon: they only record the presence of cue words (indicating greetings, answers and thanks), whether a message is a question, and whether a message is long (over ten words). Most importantly, the system does not make use of any information about message context at all.

3 Method

We first present how we adapted Elsner and Charniak's system to be used on French language data. Then, we show how we extended the base system to make use of additional discursive information.

3.1 French-language implementation

The main reason why their system requires adapting before it can be used on non-English corpora is that it relies on linguistic resources: a list of stop words, a list of technical words, and several lists of cue words to recognize greetings (3 words: "hey", "hi" and "hello"), answers (5 words: "yes", "yeah", "ok", "no" and "nope") and thanks (3 words: "thank", "thanks" and "thx").

For our adaptation, stop words (the fifty shortest words) were directly extracted from the corpus. The list of technical words was generated from different sources: some were extracted from the corpus (URLs and large numbers), and some were extracted from Ubuntu's French and English language documentations[3][4][5].

Cue words were translated into French and expanded to include more variations. Relevant English terms commonly used in French were kept. As a result we obtained a list of 24 cue words for greetings, 85 for answers and 23 for thanks.

3.2 Extension with relational information

We propose an extension to Elsner and Charniak's system that consist of the addition of new discursive information on top of existing lexical features. One of this paper's goals is to measure how discursive information can improve the performance of a chat disentanglement system.

Here, we focus on relations between messages. We present a simple annotation scheme for inter-utterance relations inspired by the DIT++ taxonomy of dialogue acts (Bunt, 2009). We distinguish between functional dependencies (such as the one between a question and an answer) and rhetorical relations (such as the one between a clarification and the utterance it relates to). Rhetorical relations are further subdivided into three classes: explicit subordination relations, explicit coordination relations and implicit coordination relations. We distinguish explicit and implicit coordination relations in order to represent differently series of utterances that are only indirectly related. For example, two consecutive questions could be merely related by the fact that they both serve to further the advancement of a common task. These situations are frequent in problem-oriented conversations such as those found in the Ubuntu platform.

These four classes allow us to represent the structure of a multi-party dialogue. There are two ways they can then be used to build features for the disentanglement system. For each message pair, we can choose to record only whether they are related in some way, or we can have a separate feature for each kind of relation. It is interesting to note that specifying the relation type can be informative and could help determine whether implicit coordination relations are relevant to the task. Because of their implicit nature we expect them to be very hard to automatically detect, so if they happen to be instrumental for chat disentanglement the overall difficulty of the task might be higher than expected.

4 Experimental framework

We first briefly describe our data, then we report our guidelines and inter-annotator agreement scores for our manual annotation tasks. Finally we present the metrics we use for evaluate the automatic disentanglement.

4.1 Corpus

The corpus was built from logs of the French language Ubuntu's IRC channel dedicated to user support[6]. It is part of an effort for building a multimodal computer-mediated communication corpus

[3]Ubuntu's French language glossary: `http://doc.ubuntu-fr.org/wiki/glossaire`

[4]Ubuntu's French language thesaurus: `http://doc.ubuntu-fr.org/thesaurus`

[5]Ubuntu's English language glossary: `https://help.ubuntu.com/community/Glossary`

[6]irc.freenode.net/ubuntu-fr

in French Hernandez and Salim (2015). The conversations found in the corpus are for the most part task-oriented. It contains 1,229 messages, all of which were manually annotated in terms of conversation and relations (See more details in Section 4.2). The corpus covers 58 different conversations, 12 of which contain only one message, for a total of 46 actual multi-participant conversations (average number of participants: 3.69). These conversations have an average length of 26 messages and a median length of 4 messages. They are interrupted on average 4 times by messages belonging to a different conversation.

4.2 Annotations and agreement

In order to compute inter-annotator agreement, 200 additional messages were annotated in terms of conversation and relation by respectively three and two annotators. Our metric is Cohen's Kappa. The annotators were French native speakers, with background in Linguistic and Natural Language Processing, but had varying levels of annotation experience.

For the conversation annotation task, we give as guidelines the following intuitive definition: Consider as a conversations, the set of utterances

- (whose content are) related to or dependent on a similar context or information need.
- and uttered by the same person or by persons in an interactive situation.

For the relation annotation task, the guidelines were based on the definition given in Section 3.2.

Both annotations tasks were carried out independently on distinct messages.

The results for the conversation annotation task, in table 1, show a very strong agreement between the three annotators.

	A_1	A_2	A_3
A_1	1.0	0.95	0.92
A_2		1.0	0.97
A_3			1.0

Table 1: Agreements for conversation annotation.

For the relation annotation task, we find an agreement of 0.80 when we consider only whether the utterances are related, and of 0.68 when we consider the particular type of each relation. These values corroborate results previously reported in the literature. Identifying a relation's correct type is a difficult task for humans; and we expect the same to be true for machines. Further work will look into the particular situations in which annotators disagree.

4.3 Metrics

We use the same metrics as Elsner and Charniak to evaluate how well different disentanglement methods perform. We use `one-to-one accuracy` to measure the global similarity between the reference and the automatic annotations. It is computed by pairing up conversations from both annotations in a way that maximizes total overlap, and then report it as a percentage. This is useful to estimate whether two annotations are globally matching or not. In contrast, the `local agreement` metric (loc_k) measures the agreement for pairs in a given context of size k. For a given message, each k previous messages are either in the same or a different conversation. The loc_k score is the average agreement of two annotators on these k pairs, averaged over all utterances. It is useful to evaluate local agreement, which is important for the analysis of ongoing conversations.

All scores are computed over 5-fold cross-validation.

5 Experiments and results

First we tried to estimate how relevant the discursive information is for recognizing conversations. Then we evaluate the adaptation of Elsner and Charniak's system for the French language, as well as the addition of discursive features.

5.1 Using discursive relations to recreate conversations algorithmically

We try to determine whether message relations are good indicators of conversational clusters. In order to do that, we project relations into conversations according to the assumption that two related messages belong to the same conversation. We obtain a new set of conversation annotations.

We then compare this new set with the reference annotations. Using the one-to-one metric to measure global similarity we find that this method performs at 0.90 accuracy. Using the loc_3 metric, we find a 0.96 agreement. This shows that discursive relations are highly valuable for the task of chat disentanglement, but also highlights the fact that there can be relations between messages of

different conversations.

5.2 Adaptation of Elsner and Charniak's system for the French language

For this experiment we compare the results of our adapted system to Elsner and Charniak's as well as the five following baselines:

- **All different:** each utterance is a separate conversation.
- **All same:** the whole transcript is a single conversation.
- **Blocks of k:** each consecutive group of k utterances is a conversation.
- **Pause of k:** each pause of k seconds or more separates two conversations.
- **Speaker:** each speaker's utterances are treated as a monologue.

Results for the adapted system are compared to each baseline in table 2. The third and fourth baselines, "blocks" and "pause", are computed with an optimal k[7].

	one-to-one	loc_3
All different	0.05	0.17
All same	0.25	0.83
Speaker	0.45	0.51
Blocks	0.49	0.83
Pause	**0.71**	0.85
System	0.68	**0.87**
System with relations	0.60	0.84

Table 2: Result comparison with each baseline.

Unlike in the experiments described by Elsner and Charniak, here the system fails to significantly outperform the "pause" baseline. It barely beats it according to the loc_k metric and is best when performance is measured by one-to-one accuracy. However, the system and the best baseline's results are both far higher than those Elsner and Charniak obtained on their corpus. Their results are reported in table 3.

This discrepancy can be explained by a structural difference between the two corpora. The loc_k metric for the "all same" baseline shows that on a local window, messages usually belong to the same conversation: simply put, our corpus is less entangled than Elsner and Charniak's.

[7]Block size is set at 105 for one-to-one accuracy and at 245 for loc_3, and pause time at 240 seconds for both metrics.

	one-to-one	loc_3
Best baseline	0.35 (Pause)	0.62 (Speaker)
System	**0.41**	**0.73**

Table 3: Results reported by Elsner and Charniak

5.3 Addition of relational discursive features

For a different experiment, we add relational features to the classifier. We choose not to consider the specific type of relation, but merely record whether two messages are related. The results are reported in table 2. We find that adding relational features in such a way do not improve the system. This may be explained by the fact that due to the way candidate pairs are selected, the system does not take message relations into account when they are separated by a certain time interval.

6 Conclusion and future work

We adapted Elsner and Charniak's disentanglement system to French and tested it on a chat corpus extracted from the French language Ubuntu platform's main IRC channel. Results were much higher than those reported in the original paper, underscoring the fact that disentangling performances are heavily correlated with how deeply conversations are intertwined in the data. The experiment also showed that a simple heuristic can be as effective as a complex trainable system when conversations are only lightly entangled. Therefore, corpus characteristics should be taken into account in order to choose an appropriate approach.

We also experimented with discursive features in the form of relational information between messages. We found that using such information to algorithmically annotate conversations yielded much more accurate results than the machine learning systems or any baseline. When we tried to use relations as feature to feed the maxent classifier, however, its global performance decreased.

Additional work is required to elaborate a typology allowing for the selection of a corpus' most appropriate disentanglement system. We also plan on performing additionnal experiments making use of the specific types of relations between messages.

Acknowledgments

This material is based upon work supported by the Unique Interministerial Fund (FUI) No. 17. It is part of the ODISAE[8] project. The authors would like to thank the anonymous reviewers for their valuable comments.

References

Harry Bunt. The DIT++ taxonomy for functional dialogue markup. In *Proceedings of the AAMAS 2009 Workshop "Towards a Standard Markup Language for Embodied Dialogue Acts" (EDAML 2009)*, pages 13–24, Budapest, Hungary, 2009.

Micha Elsner and Eugene Charniak. You talking to me? a corpus and algorithm for conversation disentanglement. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL-08)*, pages 834–842, Columbus, OH, USA, 2008.

Micha Elsner and Eugene Charniak. Disentangling chat. *Computational Linguistics*, 36(3):389–409, 2010.

Nicolas Hernandez and Soufian Salim. Construction d'un large corpus libre de conversations écrites en ligne synchrones et asynchrones en français à partir de ubuntu-fr. In *The first international research days (IRDs) on Social Media and CMC Corpora for the eHumanities*, Rennes, France, 2015.

Nam Su Kim, Lawrence Cavedon, and Timothy Baldwin. Classifying dialogue acts in multi-party live chats. In *Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation (PACLIC 2012)*, pages 463–472, Bali, Indonesia, 2012.

Elijah Mayfield, David Adamson, and Carolyn Penstein Rosé. Hierarchical conversation structure prediction in multi-party chat. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL 2012)*, pages 60–69, Seoul, South Korea, 2012.

Lidan Wang and Douglas W Oard. Context-based message expansion for disentanglement of interleaved text conversations. In *Proceedings of the 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2009)*, pages 200–208, 2009.

[8] http://www.odisae.com/