TextLink: Structuring Discourse in Multilingual Europe

STSM Report

COST STSM Reference Number: 1312-34134

Period: 2016-09-02 to 2016-09-17

COST Action: IS1312

STSM type: Regular (from Sweden to United Kingdom)

STSM Applicant: Marta Andersson, Stockholm University, Sweden; marta.andersson@english.su.se

STSM Topic: Direct and Indirect Evidence for Causal Relations

Host: Prof. Bonnie Webber, University of Edinburgh, United Kingdom; bonnie@inf.ed.ac.uk

The purpose of the current STSM was to identify and analyze lexico-syntactic evidence for forward causal relations (RESULT and PURPOSE), with a focus on the open-ended class of (possibly) multi-word indicators (alternative lexicalizations - AltLex). Several different genres were analyzed in order to tackle three main questions: (1) Can such expressions be found effectively and efficiently? (2) How frequently do they occur in different registers? (3) How reliable are such expressions at predicting the discourse relations of interest in a different register?

The answers to these questions are not straightforward, both due to the specific character of causal relations, and also because of the well-known difficulties in retrieving and categorizing AltLex. The latter phenomena, despite their functional resemblance to established discourse markers, occupy an intermediate position between signaled and implicit relations. This is particularly true of causality, which requires world knowledge and inference for retrieval and, as a consequence, is frequently left unmarked. What follows is that the choice of a relevant corpus may be quite crucial for a study of AltLex, as some genres may be naturally predisposed to signaling causality (or not).

This was observed in the course of the current project where several different text types were initially consulted for the presence and identification of causal relations. The choice of suitable corpora was, as previously declared in the proposal, the main goal of the first STSM week. To begin with, the BioDRB corpus of biological texts (Yu et al., 2008) was consulted. This corpus was previously annotated and analyzed for the presence of AltLex in all discourse relations and thus can serve as a reliable reference point in the analysis of such expressions. Subsequently, the Hansard

corpus of the Senate of Canada was consulted. Due to the commonly highly argumentative character of such speeches, this collection of texts was expected to include pragmatic causality (i.e. conclusions and claims) and, possibly, causal PURPOSE relations. Perhaps surprisingly, the findings of this part of the analysis were rather disappointing. The observations form multiple random manual searches suggest that parliamentary speakers convey causality either via explicit markers or implicit inferences. The reason for that is likely to be related to a clear pragmatic goal of this discourse type, where causality can be easily retrieved directly from the logical relations between the propositions (or, in case the clarity of the reasoning requires signaling - explicitly marked).

Subsequently, several Wikipedia weather-related texts were consulted for the same purpose. These texts were expected to include mainly causal relations of real-world RESULT. While this is generally true, the relations in the texts describing the common weather phenomena do not seem to include many causal links. More commonly, the relations conveyed are those of EXPANSION (i.e. Instantiation, List, Alternative; Prasad et al., 2008) where the author's focus is more on consecutive order of events rather than causality. This is possible because temporal order typically presupposes causal dependencies, however, causality in the weather-related Wikipedia entries is commonly marked by the verb "cause", which is a signal of forward causal relations (outside the scope of the current study). This confirms the observations that the CAUSE part of causal relation is commonly more salient for rhetorical purposes, at least in some genres (Andersson, 2016). In this particular case, the description of, for instance, blizzard is quite naturally focused on the factors causing this phenomenon, while the effect is always the same.

Those initial corpus searches have important implications with respect to questions (1) and (2). Information retrieval from unannotated corpora is laborious and prone to annotator's impaired sensitivity to textual signals present in discourse. Unannotated corpus may also result in arbitrary and subjective judgments concerning the coherence relation type. Therefore, the first important conclusion from the current STSM is that one way to increase both the reliability of the AltLex retrieved from the text (precision and recall) and also decrease the amount of texts that have to be considered, is to work with corpora annotated for discourse relations. Reducing the number of tokens to analyze should lead to increased precision and, as a result, higher recall. Another important conclusion is that, despite the omnipresent nature of causality, certain genres will not be particularly abundant in causal relations.

For these reasons, two additional corpora were consulted – a corpus of ten English TED talks and a corpus of instruction texts from the Sunset magazine. The former collection of texts is currently under annotation, and so the unannotated version had to be used. The instruction corpus

(Kim and Di Eugenio, 2006) is annotated with regard to the coherence relation type in framework based on RST (Mann and Thompson, 1988). This means that it distinguishes between backward and forward causal relations (i.e. CAUSE and RESULT), as well as PURPOSE. Both these corpora were expected to include causal relations of different kinds for different reasons. Since the goal of TED talks is to present some observations and convince the listener to the more overarching conclusion/idea that stems from these observations, this material seemed suitable both with respect to pragmatic and real-world causality. The instructions corpus was related to their specific goal-oriented character, where the texts were expected to involve both PURPOSE and RESULT relations.

Indeed, as expected, both corpora provided useful instances of AltLex. Those found in the instructions corpus are often aimed at conveying a future RESULT, i.e. PURPOSE (annotated as 'goal'). PURPOSE has not been commonly identified in the previous work as an AltLex (1 instance in BioDRB, Prasad, McRoy, Frid, Joshi and You, 2011), possibly because of its highly constrained character and overt marking requirement (Andersson and Spenader, 2014). Interestingly, the instructions corpus, includes several instances of non-prototypical marking of PURPOSE, for example:

(1) For larger or unusually shaped openings, measure the height and width of each opening.
(2) If you're planning to add a protective coating (see page 52) to a porous wall covering, always use premixed vinyl adhesive, rather than dry adhesive.

Even though a *for*-clause is a rather common means to signal PURPOSE, *for* is not an established purposive connective. In this respect it could be perceived as an alternative way to signal purposive contexts and subsumed under the 'syntactically free/lexically frozen' class proposed by Prasad, Joshi and Webber (2010). Similarly, the expression *if* could be categorized as a member of this group; however, *if* is primarily a conditional marker and will not always signal PURPOSE, which the following example from a TED talk illustrates:

(3) But *if* you consider a year's worth of emails, or maybe even a lifetime of email, collectively, this tells a lot. (TED: 2204)

Nevertheless, it is not uncommon that a connective prototypically used to signal one relation type, occurs as an AltLex of another relation. This is because causal, temporal and conditional relations commonly presuppose each other (Schmidtke-Bode, 2009). (4) below illustrates an example of *now*, which is ambiguous between an interjection function and a pragmatic RESULT relation (conclusion) marker:

(4) Well, with apologies to the Home Improvement fans, there's more growth in water than in power tools, and this company has their sights set on what they call "the new New World." That's four billion middle class people demanding food, energy and water. Now, you may be asking yourself, are these just isolated cases? (TED, 1927).

Admittedly, contexts like (4) are disputable, as their potentially resultative meaning relates to a wider context of interpretation. This means that possibly larger text chunks may serve as AltLex in some cases, including full clauses (Rysova, 2012).

With this observation in mind, the aforementioned necessity to limit the amount of text to analyze in order to efficiently search for more AltLex, becomes even more prominent. Previously discussed corpora annotated for discourse relations are one method. However, not only is this method prone to missing AlTLex (see Hidey and McKeown, 2016), but the question about accuracy of annotations arises in case of overlapping relations (e.g. (2)). This means that a range of methodologies to detect how AltLex of causality are expressed ay be needed.

Another possible method is paraphrasing unannotated corpora through back-translations (projection) (Callison-Burch, 2007; Prasad et al., 2010). This approach involves learning a recognizer to distinguish between causal and non-causal meanings of a word in one language and use it to find an alignment with phrases in another. If A has been translated as B in the target language, the alignment exists on the phrase level (possibly also on the sentence level). The advantage of looking for projection paraphrases is that no annotated corpus is needed; however, the problem with AltLex is that we know the function of some phrases, but not really the mechanism that it is being used to convey the relation. This is similar to (4) above, where the word *now* can be interpreted as resultative provided that we are able to pinpoint how causality has been expressed in that context.

One approach to this question used in computational linguistic consists in describing strings of text via identification of 'regular expressions'. This formal term could be glossed as named entities with a known causal relationship. For instance, if the nouns 'cloud' and 'rain' are found within the span of twenty words of each other, their relation is likely to be causal. The further away these entities are from each other, the less likely they will be related. Yet, if we know that certain named entities are frequently causally related, we can use them to find features and concepts that causality is often conveyed through. This way of looking for coherence relation signals also limits the size of text that has to be analyzed – once we know that two entities have a relationship, the number of relevant related concepts decreases. There are, certainly, problematic aspects of searching for

regular expressions – it may not be very precise for more abstract contexts and may also be dependent on the strength of causality link. Finally, the approach may be quite intuitive (at least at the start) and requires large corpora.

Finally, a third possible way to find AltLex more efficiently is bootstrapping. This method can be used on monolingual corpora to find new AltLex of PURPOSE and RESULT. The initianl step in this process involves inventing a classifier that would get high accuracy predicting a causal relation. In Hidey and McKeown's (2016) study, newly identified AltLex were used to search for additional causal elements. The approach yielded a significant improvement over the supervised method.

To conclude – it is still quite difficult to judge which of these methods could yield most reliable results, although there is evidence that back-translations and paraphrasing (including bootstrapping) can be effective (Callison-Burch, 2007; 2009). An additional problem is that reliability of methods can be evaluated both from the vantage point of finding relevant information and also from the point of view how the item is used as AltLex (STSM question (3)). In most cases we need to face both these questions, but finding the answers should ultimately improve detection of causal relations, as it involves categorizing elements as causal (or not). So a decision on what is sufficient evidence for a specific relation has to be made also in many vague cases. This concerns, for instance, modal verbs. Modality is very often used in pragmatic causal relations (conclusions and claims), but its polysemous character may prevent its unambiguous identification as a part of AltLex of RESULT or PURPOSE. However, AltLex has to be defined as either sufficient (stand-alone) evidence or contributory evidence. While the original claim of AltLex was based on sufficient evidence, looking at different genres, which was one of the goals of the current mission, is likely to help condition the probability that certain features are more (or less) likely to contribute to certain interpretations. It seems that searching for AlTLex requires a broader categorization and that they should not be limited only to one element. This is particularly relevant for causality, which is commonly left unmarked in the canonical sense of this term.

Thus the natural next step following the current study should be working with a large corpus. One corpus that is both annotated and involves different genres is the New York Times corpus. Not only does it include 20 million of articles but, most importantly, also metadata that allows for a comparison between different genres. As the discussion above suggests, using an annotated corpus with different genres is likely to be a reliable method to find AltLex in monolingual texts. Observations from a large corpus are likely to yield more insights into the structure of these expressions and, hopefully, into variance in the strings including synonymy, hyponymy, negation

(e.g. 'this means…' vs. 'this doesn't mean…') and so forth. Finding patterns with variations is another way of limiting searches for AltLex.

References:

Andersson, M., & Spenader, J. (2014). Result and Purpose relations with and without 'so'. *Lingua*, 148, 1-27.

Andersson, M. (2016). *The Architecture of Result relations. Corpus and experimental approaches to Result coherence relations in English*. Unpublished doctoral dissertation.

Callison-Burch, C. (2008). Syntactic constraints on paraphrases extracted from parallel corpora, *Proceedings of Conference on Empirical Methods in Natural Language Processing* (EMNLP). pages 196–205, Honolulu, October 2008.

Hidey, C., & McKeown, K. (2016). Identifying Causal Relations Using Parallel Wikipedia Articles, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1424–1433, Berlin, Germany, August 7-12, 2016.

Kim, S.N., & Di Eugenio, B. (2006). Coding Scheme Manual for instructional corpus: Identifying segments, relations and minimal units.

Mann, W. C., & Thompson, S. A. (1988). Rhetorical Structure Theory: toward a functional theory of text organization. *Text*, 8 (3), 243–281.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., & Webber, B. (2008). *The Penn Discourse Treebank 2.0*. Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008), Marrakech, Morocco. June 2008.

Prasad, R., Joshi, A. K., & Webber, B. (2010). *Realization of discourse relations by other means: Alternative Lexicalizations*. In COLING (Posters) (pp. 1023-1031), Beijing, August.

Rysova, M. (2012). Alternative Lexicalizations of Discourse Connectives in Czech, *Proceedings of the Eight International Conference on Language Resources and Evaluation* (LREC'12), 23-25 May, Istanbul, Turkey.

Schmidtke-Bode, K. (2009). *A typology of Purpose clauses*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Yu, H., Frid, N., McRoy, S., Simpson, P., Prasad, R., Lee, A., & Joshi, A. (2008). *Exploring discourse connectivity in biomedical text for text mining*. 16th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB) Linking Literature, Information and Knowledge for Biology. Toronto, July.