More than Word Cooccurrence: Exploring Support and Opposition in International Climate Negotiations with Semantic Parsing

Pablo Ruiz, Clément Plancq, Thierry Poibeau

HAL Id: hal-01326692

https://hal.archives-ouvertes.fr/hal-01326692

Submitted on 5 Jun 2016

More than Word Cooccurrence: Exploring Support and Opposition in International Climate Negotiations with Semantic Parsing

Pablo Ruiz Fabo, Clément Plancq, Thierry Poibeau

LATTICE Lab (CNRS, Ecole normale supérieure, U Paris 3,
PSL Research University, Université Sorbonne Paris Cité)
1, rue Maurice Arnoux, 92120 Montrouge, France
E-mail: {pablo.ruiz.fabo,clement.plancq,thierry.poibeau}@ens.fr

Abstract

Text analysis methods widely used in digital humanities often involve word co-occurrence, e.g. concept co-occurrence networks. These methods provide a useful corpus overview, but cannot determine the predicates that relate co-occurring concepts. Our goal was identifying propositions expressing the points supported or opposed by participants in international climate negotiations. Word co-occurrence methods were not sufficient, and an analysis based on open relation extraction had limited coverage for nominal predicates. We present a pipeline which identifies the points that different actors support and oppose, via a domain model with support/opposition predicates, and analysis rules that exploit the output of semantic role labelling, syntactic dependencies and anaphora resolution. Entity linking and keyphrase extraction are also performed on the propositions related to each actor. A user interface allows examining the main concepts in points supported or opposed by each participant, which participants agree or disagree with each other, and about which issues. The system is an example of tools that digital humanities scholars are asking for, to render rich textual information (beyond word co-occurrence) more amenable to quantitative treatment. An evaluation of the tool was satisfactory.

Keywords: climate diplomacy, semantic role labelling, digital humanities

1. Introduction

Text-analysis methods widely used in social or political sciences often involve word co-occurrence. For instance, the concept co-occurrence networks surveyed in Venturini et al. (2012), or the clustering and topic modelling approaches surveyed in Grimmer and Stewart (2013). These methods are useful in order to arrive at an overview of the content of large corpora. However, these techniques do not identify which predicates relate co-occurring elements with each other. If an actor like *France* is mentioned in the same sentence as a concept, like *stricter regulations*, which is the verb mediating between both? Is *France* in favour of, or against *stricter regulations*?

Several technologies can detect related elements in texts, and the predicate that indicates their relation. A recent approach is Open Relation Extraction (e.g. Mausam et al., 2012), where relations are identified without the need to previously specify a vocabulary of predicates or actors. The corpus we're working on consists of summaries of international climate negotiations (ENB corpus) [1]. A single sentence in this corpus can contain several support and opposition predicates, which can be verbal or nominal (see Figure 1). For this corpus, the results of a workflow based on open relation extraction tools were uneven, particularly with nominal predicates. [2] To address these challenges, we developed an application with a domain model and analysis rules which operate on outputs for semantic role labelling, syntactic dependencies and anaphora resolution, provided by a natural language processing (NLP) pipeline.

Our application identifies the points supported and opposed by actors in the negotiations, and aggregates keyphrases and DBpedia concepts [3] extracted from those negotiation points. The information is presented on an interface, providing an overview of how different actors' positions in the negotiation compare to each other.

The application helps address a current need identified by digital humanists: tools for the quantitative analysis of textual structures beyond word co-occurrences.

The paper is structured thus: Section 2 discusses related work. Section 3 describes the corpus. Section 4 presents the system. Finally, section 5 provides an evaluation and discussion. Material complementing the paper, and the system itself, are accessible from the project's website. [4]

2. Related Work

Regarding prior work on the ENB corpus [1], Venturini et al. (2014) analyzed the corpus via co-occurrence networks, using the Cortext [5] corpus cartography toolkit (Chavalarias and Cointet, 2013). However, the study does not consider which actors were linked to which concepts. Salway et al. (2014) use grammar induction to determine common actor/statement patterns in the corpus, and it could be tested whether these patterns complement our workflow's outputs.

As regards text-analysis using syntactic and semantic parsing with social sciences or humanities corpora, Diesner (2012, 2014) examined the contribution of NLP to the construction of text-based networks. Kleinnijenhuis and van Atteveldt (2014), and Van Atteveldt (2015) complement co-occurrence methods with syntactic parsing.

---

[1] The Earth Negotiations Bulletin corpus (ENB),
http://www.iisd.ca/vol12/
[2] Based on outputs from Open Information Extraction 4.0
(https://github.com/knowitall/openie) and OLLIE
(https://github.com/knowitall/ollie, Mausam et al. 2012).

[3] wiki.dbpedia.org , Auer et al. (2007). The *concepts* can sometimes refer to proper nouns or named entities.
[4] https://sites.google.com/site/climatenlp/
[5] http://docs.cortext.net

1 - Multiple verbal predicates				
The EU, with NEW ZEALAND and opposed by CHINA, MALAYSIA and BHUTAN, supported including the promotion of natural regeneration within the definitions of "afforestation" and "reforestation."				
Propositions				Predicate Type
	Actor	*Predicate*	*Negotiation Point*	
1	European_Union	supported	including the promotion of natural regeneration within the definitions of "afforestation" and "reforestation."	support
2	New_Zealand			support
3	China	~supported	including the promotion of natural regeneration within the definitions of "afforestation" and "reforestation."	opposition
4	Malaysia			opposition
5	Bhutan			opposition
2 - Nominal predicate				
Much of the discussion was on a proposal by the G-77/China to include research and development in the transport and energy sectors in the priority areas to be financed by the SCCF.				
Propositions				Predicate Type
	Actor	*Predicate*	*Negotiation Point*	
1	Group_of_77/China	proposal	to include research and development in the transport and energy sectors in the priority areas to be financed by the SCCF.	support

Figure 1: **Typical corpus sentences**. Sentence 1 has predicates *supported* and *opposed*, with several actors each. Example 2 shows a nominal predicate (*proposal*). For Sentence 1, five ‹*actor,predicate, negotiation point*› propositions are extracted by the system, and the opposing actors (*China*, *Malasia*, *Bhutan*) are assigned a proposition which is a negated version (with ~*supported* as the predicate) of the proposition for the main verb *supported*.

These studies focus mostly on syntactic dependencies and verbal predicates. We are exploring semantic role labeling as the main source of relation information, and we are addressing nominal predicates besides verbal ones. We also provide an interface to navigate the extraction results.

3. Corpus description

ENB volumes[1] are divided into issues, each of which is a 2000 word summary of the negotiations for one day in a climate summit. The corpus strives for an objective tone. To avoid biases, it uses similar syntactic structures when reporting about all participants' interventions. Typical sentences, showing how the corpus reports on participants' support and opposition in the negotiations, are in Figure 1. We analyzed the 255 issues (ca. 35,000 sentences) for the Conference of the Parties (COP) summits. We parsed the HTML into clean text, and added a date for each issue based on ENB's table of contents.

4. System architecture

The system's goal is helping researchers analyze patterns of support and opposition between different actors in climate negotiations, as well as examining which issues these actors agree or disagree about. To this end, based on the outputs of an NLP toolkit, and based on a domain model, the system applies rules to extract domain-relevant propositions, formalized as ‹*actor, predicate, negotiation point*› tuples[6] (see Figure 1). The information is made navigable, together with the original corpus, on a user interface. In addition, the system extracts keywords and linked entities from the negotiation points, and also displays them on the interface, in response to user queries.

4.1. NLP pipeline

We used the IXA Pipes NLP toolkit[7] (Agerri et al., 2014), and compatible tools. The toolkit's default English modules were used for **tokenization**, **part of speech tagging** and **constituency parsing**.

Anaphora resolution: Some types of pronominal anaphora (see 4.2) were resolved via custom rules based on coreference chains from CorefGraph[8], a Python implementation of Stanford's *dcoref* (Lee et al. 2013).

Dependency parsing and **semantic role labelling** (SRL) were carried out with ixa-pipe-srl[9], which provides a wrapper around the mate-tools library (Björkelund, Bohnet et al., 2010). The dependency and SRL format are the CoNLL ones (Surdeanu et al., 2008). SRL is performed against PropBank (Palmer et al. 2005) and NomBank (Meyers et al. 2004).

Keyphrase Extraction: We used YaTeA (Aubin and Hamon, 2006), which extracts multiple-word and single-word terms in an unsupervised manner, using syntactic and statistical criteria.

Entity Linking (EL) was performed with the ELCO3 tool from our previous work (Ruiz and Poibeau, 2015). This combines EL outputs from several public-domain EL systems, and selects the best outputs via a weighted vote.

Annotation format: IXA Pipes uses NAF, the NLP Annotation Format (Fokkens et al., 2014). This is an

---

[6] The terminology adopted is as follows: ‹*France, proposed, a new approach to carbon sinks*› is a proposition, with actor *France*, predicate *proposed* and *a new approach to carbon sinks* as the negotiation point.

[7] http://ixa2.si.ehu.es/ixa-pipes/
[8] https://bitbucket.org/Josu/corefgraph
Constituency parsing is pre-required by this tool.
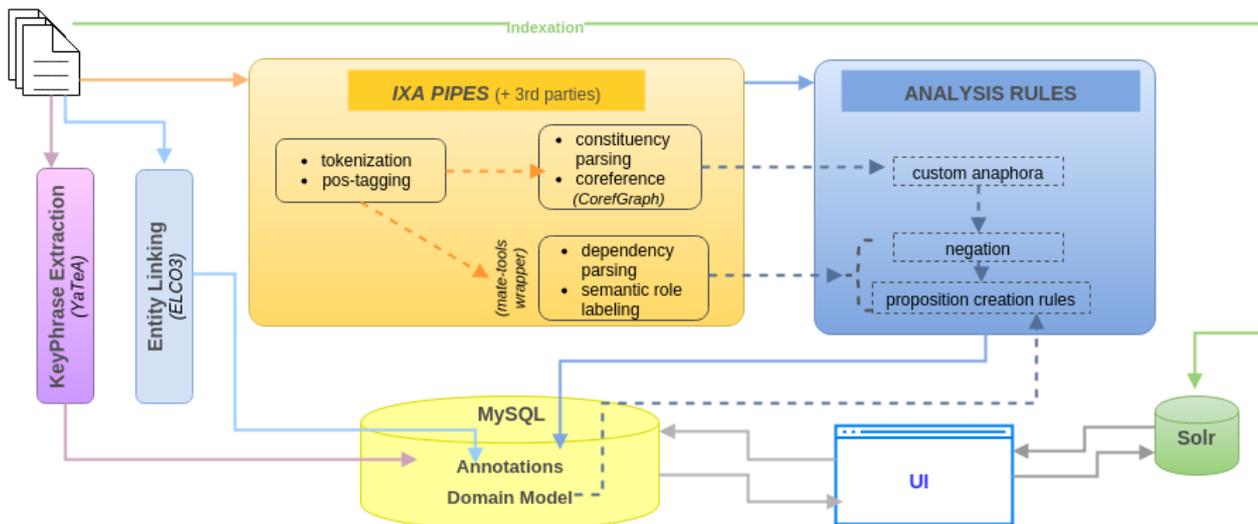[9] https://github.com/newsreader/ixa-pipe-srl

Figure 2: **System architecture**: The corpus is indexed in Solr, and enriched with different annotations, that get stored in a MySQL DB: keyphrases, linked entities, and ‹*actor, predicate, negotiation point*› propositions extracted via a domain model and analysis rules, based on the output of an NLP toolkit (IXA Pipes) providing dependencies, coreference and SRL. Users access the information on a Django-based interface.

XML format composed of layers, each of which represents an analysis step (tokenization, part-of-speech tagging, etc.). The KafNafParserPy[10] library was used to manage NAF annotations.

4.2. Domain model and analysis rules

The **domain model** contains actors and predicates. Actors represent participants in international climate negotiations,[11] and are formalized as a map between actor variants and their DBpedia URI. The model also contains lemmas for verbal and nominal predicates. Some verbs are neutral reporting verbs (e.g. *announce*). Other verbs express notions like support or opposition and agreement or disagreement (e.g. *criticize*). The verbs are contained in PropBank. The nominal predicates (e.g. *announcement, objection*) express similar notions to the verbs, and were selected from NomBank. [12] The predicate type (i.e. *support*, *oppose* or *report*) is also specified in the model.

Several **analysis rules** were implemented, that identify propositions based on a predicate's semantic roles. Most of our domain predicates involve an agent and a message (i.e. a negotiation point) expressed by that agent in a given manner: agreeing with it, objecting to it, or simply mentioning it. In that sense, actor mentions in a predicate's *A0* argument [13] correspond to the actor expressing a message, and the predicate's *A1* argument often corresponds to the negotiation point addressed by the actor. Based on this, the generic rule is in Figure 3.

Rule: Generic proposition

for each predicate p :
 · resolve negation (see below)
 for each pronoun *he, she, it* in p's A0 argument :
  · apply anaphora resolution (see below)
 for each actor-mention am in p's A0 argument :
  · create a proposition ‹$am, p, point$›, where *point* is a concatenation of p's A1 arguments

---

Figure 3: **Generic proposition** rule

In some sentences, the *A1* role contains actors rather than a negotiation point. This is notably the case in constructions like *China, opposed by the EU, preferred…* (see Figure 1 for an example). The agent of *opposed by* is the agent of a proposition that contradicts the main verb's proposition. The rule to treat such constructions is in Figure 4.

Rule: Proposition for an opposing actor

for each *opposed by* sequence ob :
 · find proposition *main* for the sentence's main verb
 for each actor-mention oam in ob :
  · create a proposition ‹$oam, \sim p_{main}, point_{main}$›, where $\sim p_{main}$ is a negated form of *main*'s predicate, and $point_{main}$ is *main*'s negotiation point

---

Figure 4: Proposition rule for **disagreeing actors**

| Docs | DBpedia | KeyPhrase |

| ActorView | ActionView | AgreeDisagree |

255 results found in 2200 ms Page 12 of 26

Actor	Action	Evidence
Canada	supported	CANADA supported the modified Option C, as did HONDURAS, who supported Iceland on including soil productivity.
Canada	urged	CANADA, supporting the GEF initiative, urged the GEF to further develop its interagency approach.
Canada	urged	CANADA, supported by SWITZERLAND and BULGARIA, urged Parties to ensure early payment of contributions.
Canada	urged	CANADA, supported by JAPAN, EU and NORWAY, urged the adoption of the good practice guidance at COP-10.
Canada	welcomed	The UK announced contributions to the work programme, and CANADA welcomed recognition of indigenous knowledge.
Canada	cautioned	CANADA cautioned that such a process could be unpredictable.
Canada	disagreed	CANADA, with the EU, disagreed that no progress had been made.
Canada	disagreed	Canada disagreed that this was the appropriate time to decide.

Nbr 18 - COP-01   More Like This

... will be a second round with three cities and, if necessary, a third round with two cities.

Estrada then noted that budget implications and the need to establish a Secretariat without doubts about administrative arrangements made a decision imperative. Canada disagreed that this was the appropriate time to decide. He said that INC-11's intent was for COP-1 to select a candidate city only if there was consensus. He urged delegates to consider using the survey later at a neutral site. Responding to a question from Iran, the Chair said the survey would be informal and confidential as a means of reaching consensus, not a binding vote.

Switzerland said it was not the time politically to pursue these consultations. He said it would also be a bad UN precedent. The decision should be made on ... ... neutral ground, possibly at the CSD meeting in New York. The US said he was persuaded by the Canadian and Swiss concerns.

Italy, supported by Poland, said he was impressed by the Chair's suggestion, and that Article 8.3 of the Convention says COP-1 should designate a Permanent Secretariat. Germany said now is an appropriate time to take a decision. Nigeria, Costa Rica, Ethiopia, Mali, Trinidad and Tobago, Togo, Cape Verde, Niger, Mauritius and Djibouti supported the Chair's proposal. Burkina Faso said that COP-1 should take the decision, but the question should be put before the Ministerial Segment.

Canada said Article 8.3 did not require COP-1 to decide on the location. He called for a substantive discussion, including budget figures comparing costs for the four locations. The Chair ... ... said the figures had been presented in consultations on the Convention budget, and participating delegations have had an opportunity to review them. The Chair said deferring the decision by a few weeks would be expensive and complicated. He said it was the feeling of the house that delegates
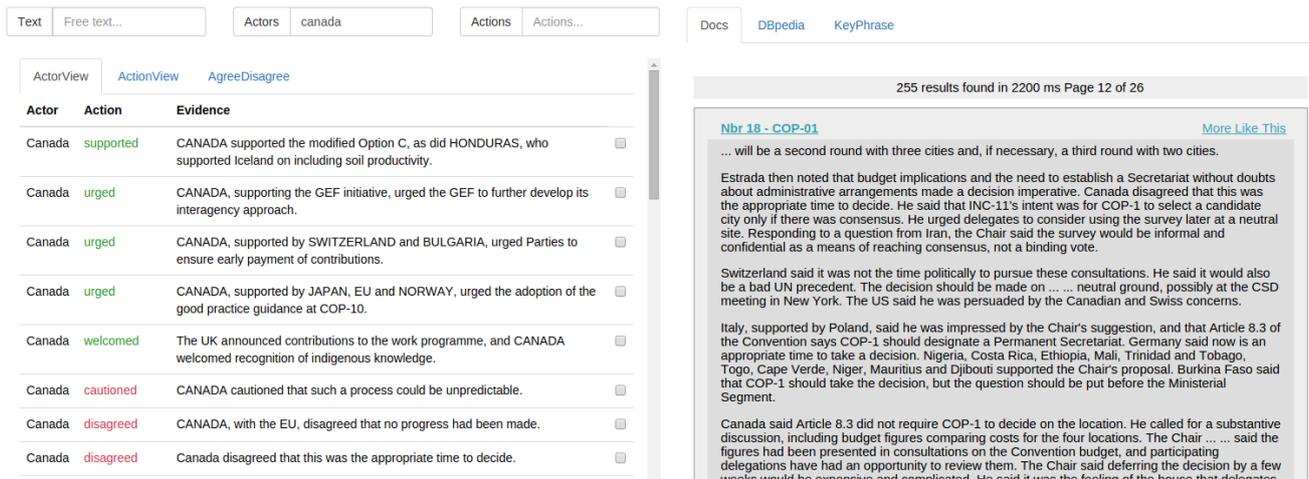
Figure 5: **Main view of the interface**. The left panel gives access to the search workflows (*Text, Actors, Actions*). It also shows propositions for a query (e.g. the actor *Canada*), and gives access to the *AgreeDisagree* view. The right panel shows the documents in the *Docs* tab, as well as aggregated keyphrases and DBpedia concepts/entities for the query or for selected propositions, in the other tabs.

Further custom analysis rules were created. A productive rule was adding *A2* roles to the proposition's negotiation point, after adding *A1*. Also, in order to make up for uncommon SRL analyses, actors were sometimes searched in adjunct roles like *AM-ADV* or *AM-MNR*.[13]

Negation is treated by finding *AM-NEG* roles related to the predicates, as well as negative lexical items (e.g. *not, lack*) in a window of two tokens preceding the predicate.

Anaphora resolution: In the corpus, a personal pronoun (*he, she*) can be the anaphor for a country[14], besides the inanimate pronoun *it*. Two rules were created to deal with this non-standard pronoun use, and anaphora resolution was limited to cases covered by these rules:

- An actor in the subject of a sentence's main verb (based on dependency parsing) is taken as the antecedent of a sentence-initial *he*/*she* in the following sentence.
- Antecedents for a pronoun (from CorefGraph's coreference chains) are only accepted if they are in the same sentence as the pronoun, or in the sentence immediately preceding the pronoun.

Finally, to facilitate searches by date-range, propositions are assigned the date of the document containing them.

4.3. User interface

The interface (Figure 5) allows researchers to explore countries' positions in the negotiation, and to compare them based on keyphrases and DBpedia concepts.

A full text search is performed with the **Text** search box. Documents matching the query are displayed on the right panel, and the propositions that have been annotated by the system in those documents are displayed on the left

panel. Propositions with a given agent or a given predicate are searched with the **Actors** and **Actions** search boxes respectively. The matching propositions are displayed on the left panel, and their documents on the right.

Tabs for **KeyPhrases** and **DBpedia** concepts on the right panel provide an overview of the content retrieved for a query. For a *Text* search, the keyphrases and concepts have been extracted from full documents. For *Actors* or *Actions* searches, the keyphrases and concepts are restricted to propositions matching the query.

The **AgreeDisagree** view (Figure 6) allows selecting two actors or groups of actors, and displays keyphrases and DBpedia concepts from propositions where those actors agree or disagree.

Exporting results or editing the model's actors and predicates is currently not allowed; this would be useful future work.

Actor 1		Actor 2		PredicateType	
European_Union	▾	China	▾	disagreement	

| Sentences | Issues |

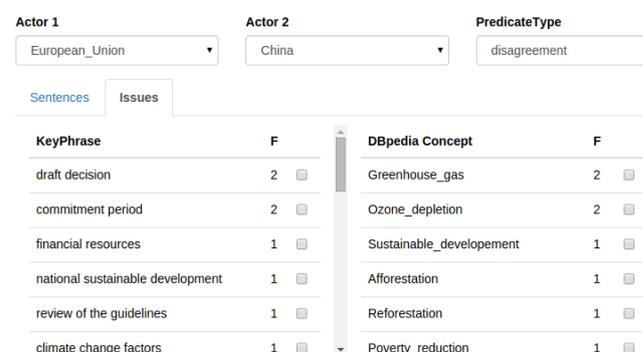| KeyPhrase	F		DBpedia Concept	F	
draft decision	2	☐	Greenhouse_gas	2	☐
commitment period	2	☐	Ozone_depletion	2	☐
financial resources	1	☐	Sustainable_developement	1	☐
national sustainable development	1	☐	Afforestation	1	☐
review of the guidelines	1	☐	Reforestation	1	☐
climate change factors	1	☐	Poverty_reduction	1	☐

Figure 6: The **AgreeDisagree View** displays keyphrases and DBpedia concepts found in sentences where selected actors (here The EU and China) agree or disagree.

---

[14] The pronoun gender corresponds to the delegate representing the country.

5. Evaluation and discussion

Several aspects can be evaluated. First, results from the NLP modules. Second, the outputs of the domain-specific components. Third, the user interface. Finally, for digital humanities applications, it is important whether the system helps researchers gain insights they would not have otherwise achieved (e.g. detect previously unseen generalizations; see Berry, 2012).

IXA Pipes modules, the keyphrase extraction and EL systems were evaluated in work already cited. Results were state of the art or competitive.[15]

The domain model and rules to create domain-relevant propositions were evaluated with two manually annotated test-sets. The first one (ENB-COP) comprises 100 sentences (311 propositions) from the COP climate summit issues in the ENB corpus, that the system was built to analyze. The test-set primarily contains sentences representing the corpus challenges (with negation, multiple actors, multiple predicates, verbal and nominal predicates).[16] The second test-set (ENB-IPCC) also comes from the ENB corpus, but not from the COP issues. Instead, it is based on ENB issues covering scientific report creation discussions by the Intergovernmental Panel on Climate Change. In other words, the ENB-IPCC test-set does not come from the same corpus that the system was developed for, and serves as a way to test the system on texts from a slightly different domain, in terms of syntactic structures and lexical items involved. The second test-set contains 283 sentences (566 propositions). For both test-corpora, a system output was considered correct if all of the proposition components (actor, predicate, negation point) match the reference exactly. Based on this notion of a correct output, precision, recall and F1 values are shown on Table 1.

Corpus	F1	P	R
ENB-COP	0.69	0.687	0.693
ENB-IPCC	0.718	0.714	0.722

Table 1: Exact-match proposition extraction. Precision, Recall, F1[17] on the ENB-COP corpus (on which system development was based), and the ENB-IPCC corpus (covering a somewhat different domain)

We consider these results acceptable for corpus exploration.

Note that our evaluation is conservative, since propositions partially matching the reference receive no credit. It could have been possible to achieve higher scores by computing F1 over individual proposition

---

[15] The results are reproduced on the project's site: https://sites.google.com/site/climatenlp

[16] Both test-sets and related information are on the project's site

[17] The definitions for these metrics were the usual:

$$F1 = 2\frac{P \cdot R}{P + R} \; ; \quad P = \frac{\text{nbr. of correct outputs}}{\text{nbr. of system outputs}} \; ; \quad R = \frac{\text{nbr. of correct outputs}}{\text{nbr. of reference outputs}}$$

---

elements, or by using the slot error rate metric (Makhoul et al., 1999). Our conservative measure avoids overestimating the system's value for our users.

The proposition elements for which the system made an error in the ENB-COP corpus are summarized in Table 2. Table 3 contains an error-type analysis in a sample containing the first 45 errors from the ENB-IPCC corpus.

Error Type	Count	% of Errors
only *predicate* wrong	2	2.1 %
only *point* wrong	63	64.95 %
both *predicate & point* wrong	32	32.99 %

Table 2: Counts and proportion of errors per error-type, for propositions of shape ‹actor, predicate, point› in the ENB-COP corpus

Error Type	Count	% of Errors
only *predicate* wrong	5	11.11 %
only *point* wrong	35	77.78 %
both *predicate & point* wrong	5	11.11 %

Table 3: Counts and proportion of errors types for propositions of shape ‹actor, predicate, point› in a sample of errors from the ENB-IPCC corpus

Most errors took place identifying the proposition's negotiation point. It can be challenging to delimit an actor's negotiation point based on semantic roles, besides the difficulty posed by our evaluation, requiring exact matches. Exploiting dependency information to add syntactically related words to the negotiation point can help (see van Atteveldt, 2015). In both test-corpora, between approx. 25% and 35% of the errors involve a wrongly identified predicate. These errors occur with some types of multi-predicate sentences.

Regarding the custom rules for pronominal anaphora, a thorough evaluation against an annotated test-set has not been performed. What can be stated based on informal evaluation is that accuracy was fine for the application's needs, but given that the rules only consider sentence initial *he/she* pronouns, coverage may be lacking.

In terms of user evaluation, a domain expert as well as two general users have provided comments. They find the application original since the data it outputs (propositions and their associated keyphrases and entities) is not available from other applications they have access to. Our users have pointed out some possible improvements. For instance, there could be more interactivity across the application's panels, e.g. clicking on a keyphrase or DBpedia entity could highlight the propositions it has been extracted from, or restrict the result set in the proposition pane accordingly. (This function is already available in the *AgreeDisagree* pane, but not elsewhere in the app). Users have also pointed out that some of the keyphrases extracted are not informative; a better weighting or filtering could be implemented.

6. Outlook

Besides improvement suggestions provided by users, mentioned above, useful future work would be exploiting the propositions extracted by the application to create network graphs representing support and opposition between negotiating parties, and between parties and issues: Opposition vs. support predicates would be represented visually by different types of edges in the network. Other useful features would be annotation export, and an annotation confidence score, which users could employ to prioritize manual result revision.

Acknowledgements

7. References

Agerri, R., Bermudez, J. and Rigau, G. (2014). IXA Pipeline: Efficient and ready to use multilingual NLP tools. In *Proceedings of LREC 2014, the 9th Language Resources and Evaluation Conference*. Reykjavik, Iceland.

Aubin, S. and Hamon, T. (2006) Improving Term Extraction with Terminological Resources. In *Advances in Natural Language Processing: 5th International Conference on NLP, FinTAL 2006*, pp. 380-387. LNAI 4139. Springer.

Auer, S., Bizer, C., Kobilarov, G., Lehman, J., Cyganiak, R., Ives, Z. (2007). DBpedia: A nucleus for a web of open data. In *The Semantic Web*, pp. 722–735. Springer.

Berry, D.M. (2012). *Understanding Digital Humanities*, pp. 1–20. Palgrave Macmillan.

Björkelund, A., Bohnet, B., Hafdell, L., and Nugues, P. (2010). A high-performance syntactic and semantic dependency parser. In *Coling 2010: Demonstration Volume*, pp. 33–36, Beijing.

Chavalarias, D. and Cointet, J.P. (2013). Phylomemetic patterns in science evolution—the rise and fall of scientific fields. *PloS one*, 8(2)

Diesner, J. (2012). *Uncovering and managing the impact of methodological choices for the computational construction of socio-technical networks from texts*. PhD Thesis. Carnegie Mellon University.

Diesner, J. (2014). ConText: Software for the Integrated Analysis of Text Data and Network Data. In *Social and Semantic Networks in Communication Research*, at *ICA, Conference of International Communication Association*.

Fokkens, A., Soroa, A., Beloki, Z., Ockeloen, N., Rigau, G., van Hage, W.R and Vossen, P. (2014). NAF and GAF: Linking Linguistic Annotations. In *Proceedings 10th Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation*. Reykjavik, Iceland.

Grimmer, J., and Stewart, B.M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, pp. 1–31. OUP.

Kleinnijenhuis, J. and van Atteveldt, W. (2014). Positions of parties and political cleavages between parties in texts. In Kaal et al (eds.). *From Text to Political Positions, Text analysis across disciplines,* pp. 1–18. Benjamins.

Lee, H., Chang, A., Perisman, Y., Chambers, N., Surdeanu, M. and Jurafsky, D. (2013). Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics 39*(4).

Makhoul, F.K., Schwartz, R., and Weischedel, R. (1999). Performance measures for information extraction. In *Proceedings of DARPA broadcast news workshop*, pp. 249–252.

Mausam, Schmitz, M., Bart, R., Soderland, S., and Etzioni, O. (2012). Open language learning for information extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 523–534.

Meyers, A., Reeves, R., Macleod, C., Szekely, R., Zielinska, V., Young, B., and Grishman, R. (2004). The NomBank project: An interim report. In *HLT-NAACL 2004 workshop: Frontiers in corpus annotation*, pp. 24–31.

Palmer, M., Gildea, D., Kingsbury, P. (2005). The Proposition Bank: A Corpus Annotated with Semantic Roles. *Computational Linguistics Journal*, 31:1.

Ruiz, P. and Poibeau, T. (2015). Combining Open Source Annotators for Entity Linking through Weighted Voting. In *Proceedings of *SEM. Fourth Joint Conference on Lexical and Computational Semantics,* pp 211–215. Denver, U.S.

Salway, A., Toulieb, S and Tvinnereim, E. (2014). Inducing information structures for data-driven text-analysis. *Proceedings of the ACL Workshop on Language Technologies and Computational Social Science*, pp 28–32.

Surdeanu, M., Johansson, R., Meyers, A., Màrquez, L. and Nivre, J. (2008). The CoNLL-2008 shared task on joint parsing of syntactic and semantic dependencies. In *Proceedings of the Twelfth Conference on Computational Natural Language Learning*, pp. 159–177. Association for Computational Linguistics.

Van Atteveldt, W., Sheaferm, T., Shenhav, S. and Fogel-Dror, Y. (2015). Clause analysis: using syntactic information to enrich frequency-based automatic content analysis. In *Symposium New Frontiers of Automated Content Analysis in the Social Sciences*, at the University of Zurich.

Venturini, T. and Guido, D. 2012. Once upon a text: an ANT tale in Text Analytics. *Sociologica, 3*:1–17. Il Mulino, Bologna.

Venturini, T., Baya Laffite, N., Cointet, J.P., Gray, I., Zabban, V. and Kari De Pryck. (2014). Three maps and three misunderstandings: A digital mapping of climate diplomacy. *Big Data & Society 1* (2), pp 1–19.