Which aspects of discourse relations are hard to learn?

Primitive decomposition for discourse relation classification

Charlotte Roze[1] Chloé Braud[1] Philippe Muller[2]

[1]LORIA, CNRS, Université de Lorraine – Nancy, France

[2]IRIT, CNRS, Université de Toulouse – Toulouse, France

September 13, 2019

Discourse Relations Identification

- discourse parsing: identification of discourse structure
  - semantic and pragmatic links between discourse units (text spans: clauses, sentences, paragraphs)
- discourse relations: explicit or implicit

(1) Climate change is caused by anthropic activities,
but politics are not doing anything about it.
Comparison.Concession.Contra-expectation (PDTB label)

(2) Climate is changing.
Humans generate too much CO_2.
Contingency.Cause.Reason (PDTB label)

Discourse Relations Identification: Difficulties

- several theories or frameworks for representing discourse structure:
  - *Rhetorical Structure Theory* (Mann and Thompson, 1988)
  - *Segmented Discourse Representation Theory* (Asher and Lascarides, 2003)
  - *Penn Discourse TreeBank* (Prasad et al., 2007)

→ corpora annotated following these various frameworks

- **no consensus** on the label sets of discourse relations
  - ± specific relations (various levels of **granularity**)

(SDRT)		(RST)
		Antithesis
Contrast	⟷	*Concession*
		Contrast

Discourse Relations Identification: Difficulties

- BUT common range of semantic and pragmatic information
- find a way to represent this common information?

Discourse Relations Identification: Difficulties

- classification task: explicit/implicit relations

- implicit relations classification: hardest task
  - "low" results (up to 51% in F_1 for less specified relations from PDTB)
  - despite the variety of approaches that have been tried

Is the problem only about data representation
or also about the **way we model the task**?

Decompose Relations into Primitives

Act on the way we model the task:

- split it into **several simpler tasks**
  - → decompose the problem
  - → investigate reasons of difficulties in discourse relations identification
- decompose information encoded by relation labels into values for a small set of characteristics: **primitives**

Decompose Relations into Primitives

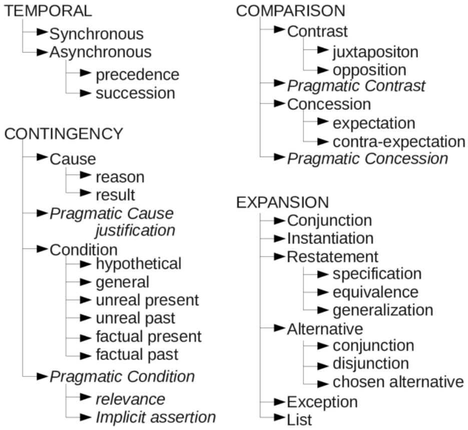
Cognitive Approach to Coherence Relations (CCR)

- inventory of **cognitively motivated** dimensions (primitives) of relations (Sanders et al., 2018)
- mappings from PDTB (2.0), RST, SDRT relations into primitives values
  - core primitives: original CCR (Sanders et al., 1992, 1993) primitives
  - additional primitives: introduced to explicit specificities of the various frameworks
- **interface** between existing frameworks

Approach

1. Operational mapping
   - annotated relations → sets of primitives values
   - tested on PDTB 2.0
2. Which primitives are harder to predict?
   - classification task for each primitive
3. Reverse mapping
   - set of primitives values → compatible relation labels
   - relation identification system

PDTB's hierarchy



- 3 levels hierarchy ($\neq$ granularities)
  - $\pm$ specific relations
  - classes, types, sub-types
- end-labels
  $\rightarrow$ most specific relations (level 3 or level 2)
- intermediate labels
  $\rightarrow$ **underspecified** relations

Primitives

PDTB relation → set of primitive values

Primitives

PDTB relation → set of primitive values

- 5 core primitives → 2 or 3 values
    - *polarity*
    - *basic operation*
    - *source of coherence*
    - *implication order*
    - *temporal order*

\+ NS (non-specified): ambiguities

- several possible values in CCR mapping
- intermediate labels ($\notin$ CCR mapping)

Primitives

PDTB relation $\rightarrow$ set of primitive values

- 5 core primitives $\rightarrow$ 2 or 3 values + NS (non-specified): ambiguities
  - *polarity*
  - *basic operation*
  - *source of coherence*
  - *implication order*
  - *temporal order*
  - several possible values in CCR mapping
  - intermediate labels ($\notin$ CCR mapping)

- 3 additional primitives $\rightarrow$ binary ($-$ or $+$)
  - *conditional*
  - *alternative*
  - *specificity*

Mapping to Primitives

Illustrate mapping into core primitives

Comparison.Concession.Contra-expectation

(3) a. The biofuel is more expensive to produce, (P)
b. **but** by reducing the tax the government makes it possible to sell the fuel for the same price. (not-Q)

- expected **implication** (P $\rightarrow$ Q): the biofuel costs more (Q)
- **denial** of this expectation: the biofuel doesn't cost more (not-Q)

Mapping to Primitives

Relation	Basic op.	Pol.	Impl. order	SoC	Temp.
Contra-expectation	cau	neg	basic	NS	NS

- involves an **implication**: *basic operation* = *causal*
  - otherwise *additive*
- involves a **negation**: *polarity* = *negative*
  - otherwise *positive*
- premise of implication in first argument: *implication order* = *basic*
  - *non-basic* (conclusion in first argument)
  - NA (non-applicable) for *additive* relations

Mapping to Primitives

- *source of coherence*: common distinction (RST)
  - *objective*: level of **propositional content**
  - *subjective*: **epistemic/speech act** level

Contingency.Pragmatic cause.Justification

(4) a. (I say that) Mrs Yeargin is lying.
b. (because) They found students (...) who said she gave them similar help.

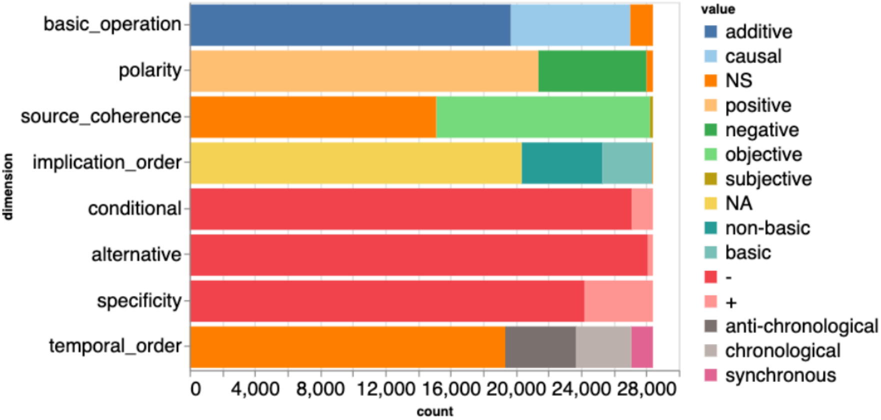
Relation	Basic op.	Pol.	Impl. order	SoC	Temp.
Justification	cau	pos	non-b	sub	NS

- *temporal order*: *chronological*, *anti-chronological*, *synchronous*

With respect to PDTB hierarchy, primitives are not of equal importance

- able to make distinctions between top-level classes (level 1)
  - *basic operation*
    - *Contingency* class → value *causal*
    - *Temporal* class → value *additive*
  - *polarity*
    - *Comparison* class → value *negative*
    - *Contingency* and *Temporal* classes → value *positive*
- label distinctions at level 2 (*source of coherence*) or 3 (*implication order*)

- mapping applied to each relation in PDTB 2.0
- 2,159 articles from the Wall Street Journal
- distribution of values for each primitive:



Experimental Setting

- classification task for each primitive independently
- training set: 28,402 pairs of arguments

Model architecture

- Each argument representation: Infersent sentence encoder (very common for semantic tasks)
  - pretrained word embeddings (GloVe)
  - encoded with a bi-LSTM with max pooling (dimension: 1024)
- Combination of the 2 arguments representations (dimension: 4096)
  - concatenation
  - absolute difference
  - element-wise product

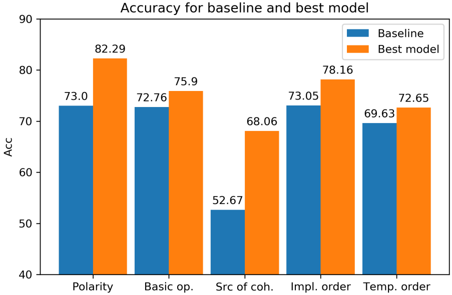
Experimental Setting

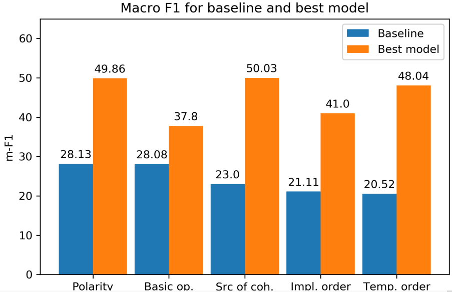
Hyper-parameters

- maximum 15 epochs and early stopping
- size of hidden layer: 0 (no layer), 512, or 4096
- regularization values: 10^{-n} with $n \in \{-8, 1\}$

Compare results (test set: section 23)

- Best model: best setting on the development set
- Baseline: majority classifier





- for 33% of argument pairs all primitives are correctly predicted
- in average, 82% primitives are correctly predicted (between 6 and 7 primitives on a total of 8)

- *polarity* and *basic operation*:
    - most “important” primitives
    - similar distribution of values

- *polarity* and *basic operation*:
  - most “important” primitives
  - similar distribution of values
- *basic operation*: lowest improvement (on all measures) wrt. baseline
  - only 17% *causal* relations correctly labeled (relations are mainly labeled as *additive*)

- *polarity* and *basic operation*:
    - most “important” primitives
    - similar distribution of values
- *basic operation*: lowest improvement (on all measures) wrt. baseline
    - only 17% *causal* relations correctly labeled (relations are mainly labeled as *additive*)
- better results for *polarity* (greater improvement wrt. baseline)
    - 50% *negative* relations correctly labeled (95% of *positive* relations)

- *polarity* and *basic operation*:
  - most “important” primitives
  - similar distribution of values
- *basic operation*: lowest improvement (on all measures) wrt. baseline
  - only 17% *causal* relations correctly labeled (relations are mainly labeled as *additive*)
- better results for *polarity* (greater improvement wrt. baseline)
  - 50% *negative* relations correctly labeled (95% of *positive* relations)
- *source of coherence*: greatest improvement wrt. baseline, but this result must be tempered
  - very small number of *subjective* relations in our dataset (less than 1%)
  - $\simeq$ only *objective* and NS values (not so much information)

- *polarity* and *basic operation*:
  - most “important” primitives
  - similar distribution of values
- *basic operation*: lowest improvement (on all measures) wrt. baseline
  - only 17% *causal* relations correctly labeled (relations are mainly labeled as *additive*)
- better results for *polarity* (greater improvement wrt. baseline)
  - 50% *negative* relations correctly labeled (95% of *positive* relations)
- *source of coherence*: greatest improvement wrt. baseline, but this result must be tempered
  - very small number of *subjective* relations in our dataset (less than 1%)
  - $\simeq$ only *objective* and NS values (not so much information)
- *temporal order*: low improvement wrt. baseline (on accuracy)
  - relations are mainly labeled as NS (majority class)

Reverse mapping

- performance of our systems on predicting discourse relations
  - reverse mapping: set of predicted values for each primitive
    $\rightarrow$ set of **compatible relation labels**

Reverse mapping

- performance of our systems on predicting discourse relations
  - reverse mapping: set of predicted values for each primitive
    $\rightarrow$ set of **compatible relation labels**

1. set containing all possible relations (at any level of the hierarchy)

Reverse mapping

- performance of our systems on predicting discourse relations
  - reverse mapping: set of predicted values for each primitive
    → set of **compatible relation labels**

1. set containing all possible relations (at any level of the hierarchy)
2. remove relations **incompatible** with the primitive values predicted
   - *polarity* is *positive*

   ⇒ all relations associated with *negative polarity* are excluded

   (same for each primitive)

Reverse mapping

- performance of our systems on predicting discourse relations
  - reverse mapping: set of predicted values for each primitive
    → set of **compatible relation labels**

1. set containing all possible relations (at any level of the hierarchy)
2. remove relations **incompatible** with the primitive values predicted
   - *polarity* is *positive*
   - ⇒ all relations associated with *negative polarity* are excluded

   (same for each primitive)
3. remove redundant information
   - set contains all sub-types of a type (or all types under a class)
   - ⇒ only keep the upper lever underspecified relation (type or class)
   - *Temporal*, ~~*Temporal.Asynchronous*~~, ~~*Temporal.Synchrony*~~

Evaluation

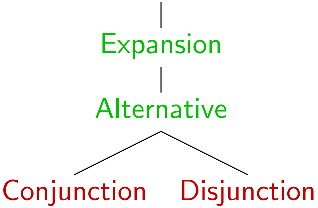
Our approach raises a number of questions with respect to the evaluation

- measure for hierarchical classification
  - underspecifications (predicted label $\pm$ specific than gold label)
- measure for multi-label classification
  - disjunction of relations (reverse mapping: set of possible relations and multiple relations in PDTB)

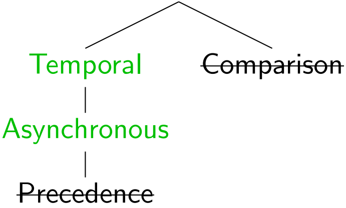
- hierarchical precision and recall (Kiritchenko et al., 2005)
- on the set of labels (at all levels)

Evaluation

- gold: *Expansion.Alternative*
- predicted:
  Expansion.Alternative.Conjunction
  Expansion.Alternative.Disjunction
- Recall = 1, Precision = 0.5



- gold:
  Temporal.Asynchronous.Precedence
  Comparison
- predicted: *Temporal.Asynchronous*
- Recall = 0.5, Precision = 1



Evaluation

Compare the performance of 2 systems (hierarchical scores)

- [Primitives] $\rightarrow$ 1 or more relations
  reverse mapping between predicted primitives to compatible relations
- [Relations] $\rightarrow$ 1 relation
  direct discourse relations classification (no decomposition)

Measures

- accuracy
- hierarchical precision and recall (**h-R** & **h-P**)
- hierarchical scores only on best match predicted relations/PDTB relations (**max-h-R** & **max-h-P**)

	Acc	h-R	h-P	max-h-R	max-h-P
			All		
Baseline	20.03	27.65	29.97	28.97	30.98
Primitives	34.15	28.89	19.32	49.07	**59.05**
Relations	**45.35**	**52.97**	**54.95**	**55.42**	56.58
			Explicit		
Baseline	23.5	25.35	26.13	27.02	27.33
Primitives	46.27	35.56	26.43	59.93	**69.59**
Relations	**59.08**	**63.63**	**65.3**	**67.4**	67.8
			Implicit		
Baseline	15.73	30.5	34.72	31.38	35.5
Primitives	19.12	20.63	10.52	35.61	**45.99**
Relations	**28.35**	**39.76**	**42.11**	**40.57**	42.67

- **missing** a lot of *Contingency* class relations (83%)
  - consistent with results on primitive prediction:
    missing value *causal* for primitive *basic operation*
  - ⇒ plain error but only one primitive is wrong in many cases

- **wrongly** predicting *Temporal* class relations (86%)
  - associated with underspecified values for primitives (kind of default relation)

- predicting primitives leaves too much underspecification (impact on recall)
- predicting too many labels (impact on precision)

Conclusion and perspectives

- one of the most important primitives (*basic operation*) seems to be hardest to predict
- primitives are **not independent** from each other
  - learning them independently < learning fully specified relation
  - future work: **multi-task** learning setting
- Extend the approach: apply this decomposition to other discourse frameworks (RST or SDRT)
  - **cross-corpora** training and prediction

Thank you!

Nicholas Asher and Alex Lascarides. Logics of conversation. Cambridge University Press, 2003.

Svetlana Kiritchenko, Stan Matwin, and A. Fazel Famili. Functional annotation of genes using hierarchical text categorization. In in Proc. of the BioLINK SIG: Linking Literature, Information and Knowledge for Biology (held at ISMB-05), 2005.

William C. Mann and Sandra A. Thompson. Rhetorical Structure Theory: Toward a functional theory of text organization. Text, 8:243–281, 1988.

Rashmi Prasad, Eleni Miltsakaki, Nikhil Dinesh, Alan Lee, Aravind Joshi, Livio Robaldo, and Bonnie Webber. The Penn Discourse TreeBank 2.0 annotation manual, 2007.

Ted Sanders, Wilbert Spooren, and Leo Noordman. Toward a taxonomy of coherence relations. Discourse Processes, 15:1–35, 01 1992. doi: 10.1080/01638539209544800.

Ted Sanders, Wilbert Spooren, and Leo Noordman. Coherence relations in a cognitive theory of discourse representation. Cognitive Linguistics, 4:93–134, 01 1993. doi: 10.1515/cogl.1993.4.2.93.

Ted J.M. Sanders, Vera Demberg, Jet Hoek, Merel C.J. Scholman, Fatemeh Torabi Asr, Sandrine Zufferey, and Jacqueline Evers-Vermeul. Unifying dimensions in coherence relations: How various annotation frameworks are related. Corpus Linguistics and Linguistic Theory, 2018. doi: 10.1515/cllt-2016-0078. URL `http://www.degruyter.com/downloadpdf/j/cllt.ahead-of-print/cllt-2016-0078/cllt-2016-0078.xml`.

Results for all primitives

Primitive	Best model		(Gain over baseline)
	Acc	m-F_1	w-F_1
Basic op.	75.90 (+3.14)	37.80 (+9.72)	69.03 (+7.74)
Polarity	82.29 (+9.29)	49.86 (+21.73)	80.59 (+18.99)
Src of Coh.	68.06 (+15.39)	50.03 (+27.03)	67.44 (+31.10)
Impl. order	78.16 (+5.11)	41.00 (+19.89)	74.89 (+13.21)
Temp.	72.65 (+3.02)	48.04 (+27.52)	69.32 (+12.16)
Cond.	98.55 (+2.67)	–	–
Altern.	98.84 (+0.06)	–	–
Specif.	85.13 (+2.20)	–	–

Results to add and other perspectives

- Score for predicting all primitives together
- Results for primitive prediction on explicit/implicit
- Distribution of explicit/implicit by relation/primitive values?

- Which models perform better on which primitive?
- Work on separate tasks, with more specific data (and more data), in order to improve the global task?
- When learning primitives on a training corpus without some relations, can we predict them correctly based on their conceptual decomposition?

		Explicit			Implicit		
		acc	w-f1	m-f1	acc	w-f1	m-f1
Basic op.	baseline	**73.14**	61.79	**28.16**	**72.3**	60.68	**27.97**
	primitives	**77.96**	72.4	**42.42**	**73.34**	64.07	**32.01**
Polarity	baseline	**66.95**	53.69	**26.73**	**80.49**	71.8	**29.73**
	primitives	**84.16**	83.4	**54.66**	**79.97**	73.67	**33.56**
SoC.	baseline	37.46	20.42	18.17	56.31	40.57	24.02
	primitives	75.24	75.11	55.71	59.17	56.17	36.38
Impl. order	baseline	73.35	62.07	21.16	72.69	61.2	28.06
	primitives	83.11	81.38	49.07	72.04	65.92	39.61
Temp. order	baseline	68.0	55.04	20.24	71.65	59.82	20.87
	primitives	75.97	73.68	54.77	68.53	63.66	30.29
Conditional	baseline	92.55	–	–	100.0	–	–
	primitives	97.59	–	–	99.74	–	–
Alternative	baseline	99.37	–	–	98.05	–	–
	primitives	99.48	–	–	98.05	–	– -
Specificity	baseline	96.64	–	–	65.93	–	–
	primitives	96.85	–	–	70.61	–	–

Test set: 1722

Nb relations	Contingency	Comparison	Temporal	Expansion
Gold	430	440	208	725
Primitives	86	296	1341	1342
Relations	467	259	120	876