Content and Annotation Compatibility Across Corpora: The Case of Events

Chantal van Son, Oana Inel, Lora Aroyo, Roser Morante and Piek Vossen

{c.m.van.son, oana.inel, lora.aroyo, r.morantevallejo, piek.vossen}@vu.nl

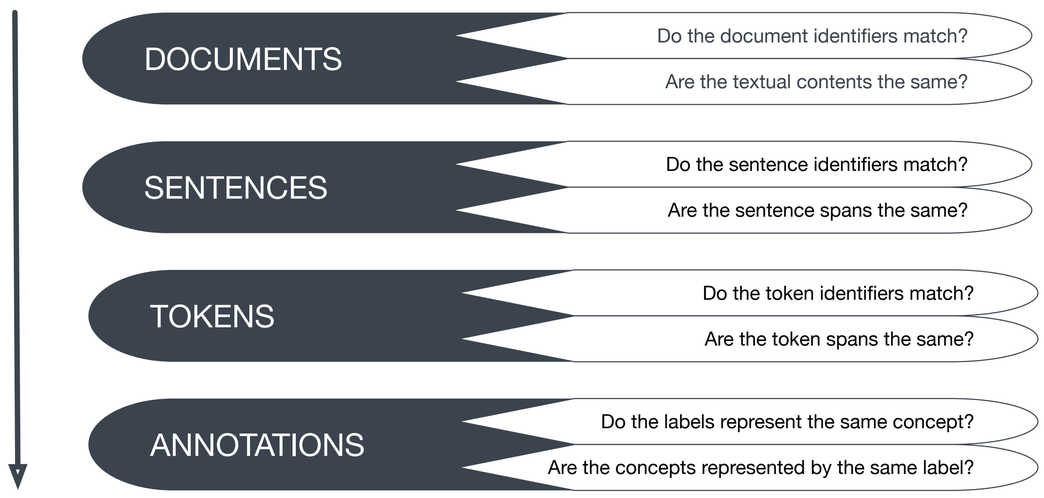
VU UNIVERSITY AMSTERDAM

Opportunities

- Many corpora available that relate to each other in terms of content (i.e. documents, sentences, tokens) and annotations
- Much to learn and gain from these corpus networks, especially if corpora (partially) overlap in both content and annotations:
  - enriching corpora by merging complementary annotations
  - comparing annotations from similar frameworks to reveal differences in conceptualization

Challenges: Content & Annotation Compatibility

However, problems with respect to alignment, comparison and merging tend to accumulate rapidly due to differences in several fundamental information layers:



Measuring Content & Annotation Compatibility

Two ways to measure compatibility:

- **Token-based comparison:** analyse which tokens are shared and if they are annotated in the same way or differently (requires content compatibility)
- **Type-based comparison:** derive statistics on the types (e.g. lemmas or POS tags) annotated with a certain label and compare the distributions across corpora, which will provide a more general insight into annotation compatibility

		PROPBANK/NOMBANK			FACTBANK		
token	POS	sent_id	token_id	annotation	sent_id	token_id	annotation
maker	NN	0	6	PRED.	4	6	-
announced	VB	0	8	PRED.	4	8	EVENT
effectiveness	NN	0	10	PRED.	4	10	EVENT
registration	NN	0	13	PRED.	4	13	-
statement	NN	0	14	PRED.	4	14	-
debentures	NN	0	26	PRED.	4	25	-
due	JJ	0	27	-	4	26	EVENT
said	VB	1	2	PRED.	5	2	EVENT
issued	VB	1	8	PRED.	5	7	EVENT
price	NN	1	13	PRED.	5	11	-
convertible	JJ	1	27	-	5	23	EVENT
maturity	NN	1	33	-	5	29	EVENT
price	NN	1	37	PRED.	5	33	-

Aligned tokens annotated as event/predicate

POS	PropBank/NomBank	FactBank	TempEval-3
VB	100.00	62.72	62.55
NN	100.00	10.51	7.90
JJ	0.00	4.77	2.77
CD	0.00	7.55	0.00
SYM:$	0.00	25.96	0.00
SYM:#	0.00	13.04	0.00
RP	0.00	1.55	0.00
RB	0.00	1.28	0.41
IN	0.00	0.61	0.16
CC	0.00	0.17	0.00
MD	0.00	0.14	0.10
DT	0.00	0.13	0.00

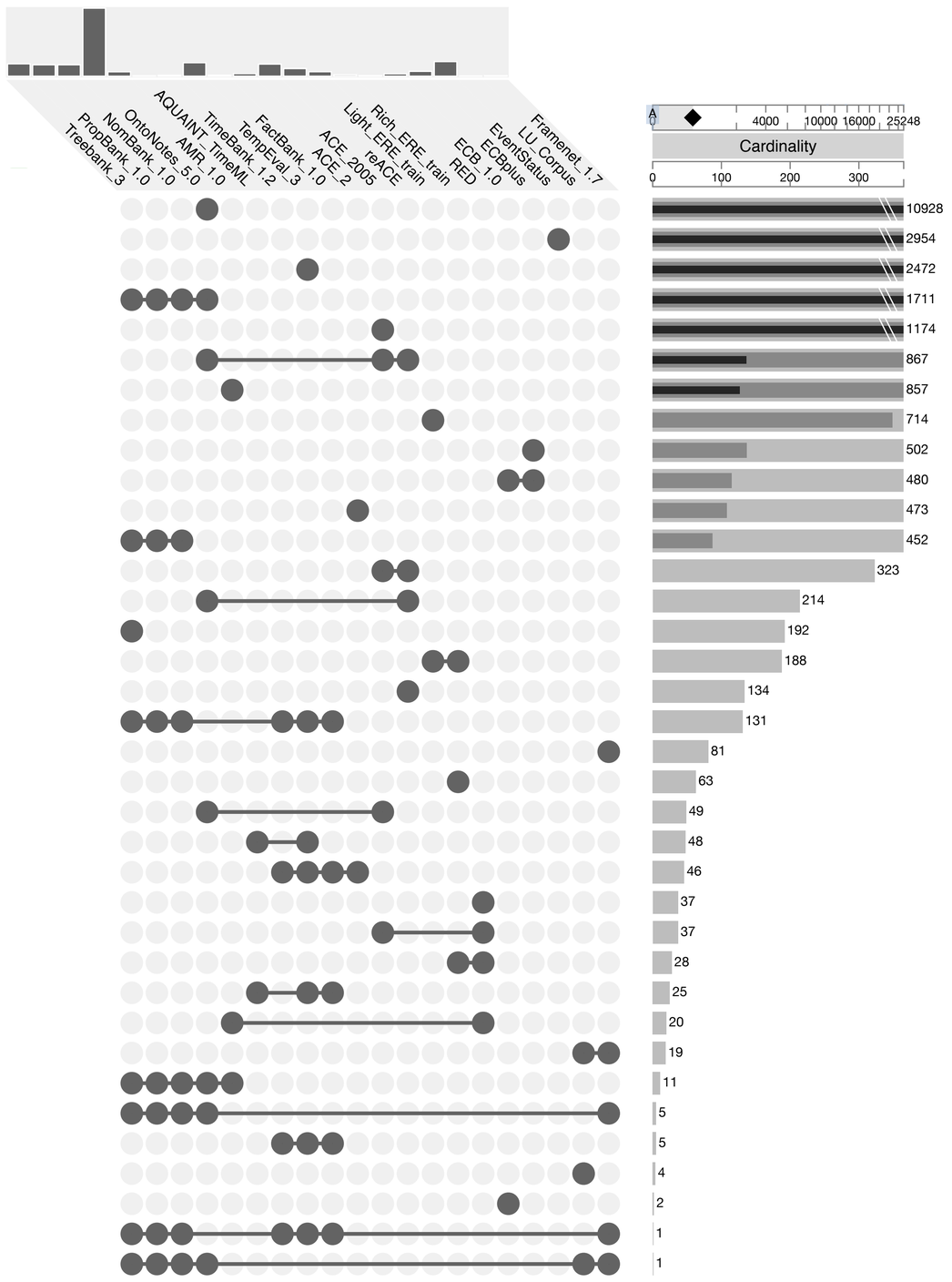
Percentage of POS-tags annotated as event/predicate

Use case: Assume we want to enrich FactBank's event factuality annotations with semantic role annotations in PropBank/NomBank (sharing 132 documents). How compatible are they in terms of content and annotations?

Data: Event-Annotated Corpora

- We selected 20 corpora that are connected to each other in terms of annotations and/or content:
  - annotated with events (or predicates), or;
  - share documents with the other corpora (and thus contain event-annotated documents)

The figure below visualizes the document intersections of these corpora on the basis of mapped document identifiers and reveals the complex network of corpora that is the result of more than a decade of data selection and extension.



This visualisation was created using Upset (http://caleydo.org/tools/upset).
For an interactive version, see https://github.com/ChantalvanSon/CorpusComparison.

Conclusion

- Possibly many opportunities for comparing and/or merging similar or complementary annotations in corpora that overlap in content
- Opportunities hampered by wide range of interoperability issues
- We propose using token-based comparison and type-based comparison for assessing compatibility
- We hope to make the community aware of these challenges and opportunities and to inspire finding ways to collaboratively improve resource interoperability

Amsterdam Data Science

This work was supported by the Amsterdam Academic Alliance Data Science (AAA-DS) Program Award to the UvA and VU Universities, by the Dutch National Science Foundation from the Spinoza-2013 price and by the European Union's 7th Framework Programme via the NewsReader project (ICT-316404).