Argumentation in Text: Discourse Structure Matters

18. 5. 2018

Lucie Findejsová

Obsah

Detekce argumentace v textu

Binární klasifikační úkol: *positive class* × *negative class*

Communicative Discourse Tree (CDT)

New Intense Argumentation dataset (NIAD)

Úvod

Argument jako klíčový bod textu

Cíl: rozpoznání vlastností diskurzu v textu

Systematické vybrání vzorů argumentace

Sestavení *New Intense Argumentation Dataset*

Porovnání výkonu několika metod učení

Communicative Discourse Tree (CDT)

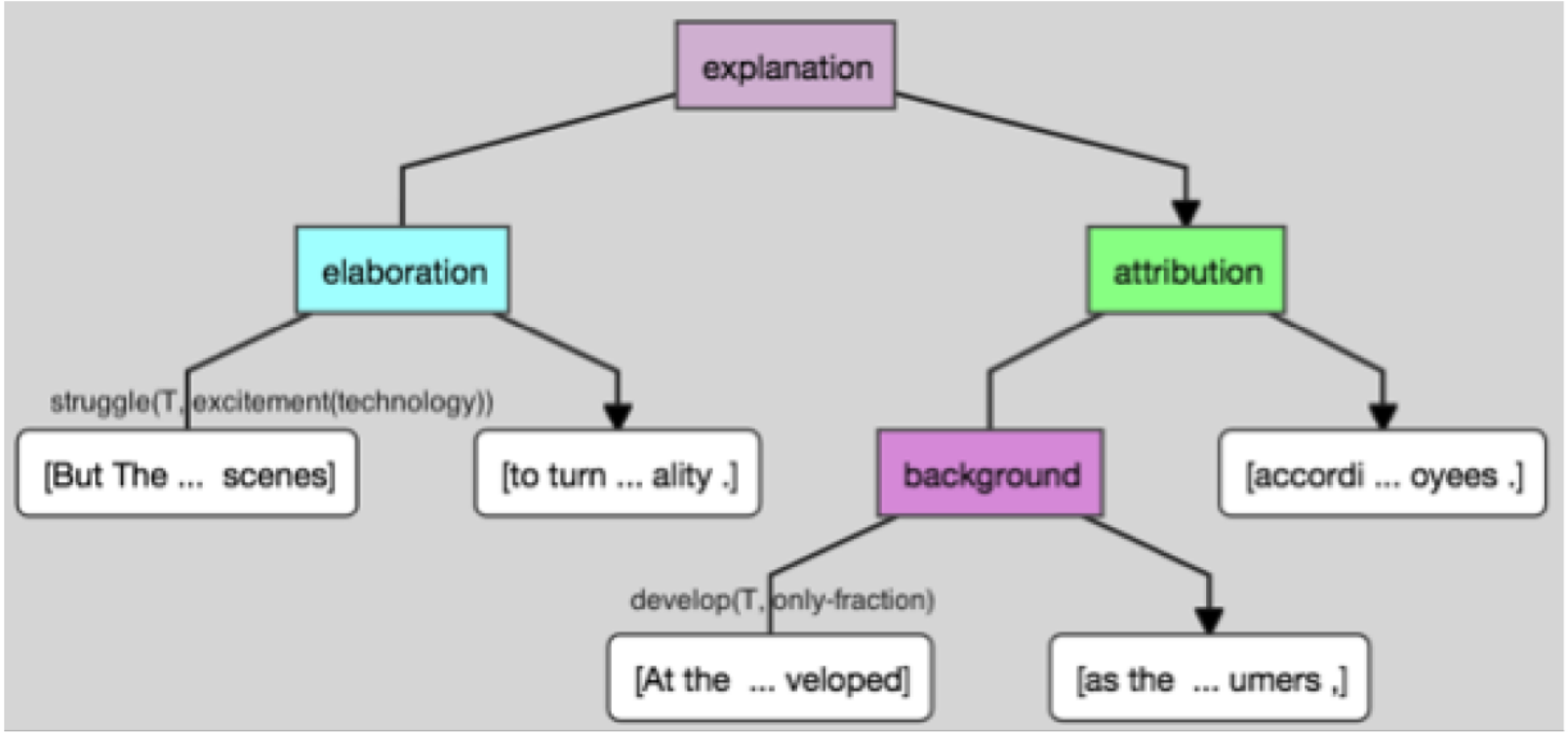
CDT – formální struktura, klasifikace textu, predikce štítku

- DT (Discourse Tree) – založen na rétorických vztazích, vybraných z textu podle *Rhetoric Structure Theory*

  má štítky na hranách grafu, což jsou komunikační akce
- CA (Communicative Actions) – bere formu slovesa, druh interakce

  (vysvětlit, konstatovat, připomenout, nesouhlasit, popřít, …)

Communicative Discourse Tree (CDT)



Nastavení klasifikace textu

Dva typy učení na CDT grafové reprezentaci

- 1) *Nearest Neighbour (kNN)* – podobnost mezi grafovou reprezentací a elementem z trénovacího setu
- 2) *SVM Tree Kernel* – přidává informaci o komunikačních akcích

New Intense Argumentation Dasaset

Sesbírání argumentačních textů

Validní × nevalidní argumentace

- Nevalidní: nesouhlas se vším, upřednostňování
  - Nepravdivá stížnost = většinou nevalidní
  - Nepravdivá stížnost + validní argumentační vzory = ?
- Validní: faktické zkreslení, neplatné sliby
  - Slabá/chybná argumentace = nevalidní

Vnitřní anotátorská shoda přes 80 %

„Intense“ kvůli značné emocionalitě argumentů

Additional Evaluation Dataset

Součást je NIAD, style & genre recognition dataset

Dvě části: positive × negative

Positive: jakýkoli druh argumentace, různé styly, žánry, typy

- *The New York Times* – 1400 článků, *The Boston Globe* – 1150 článků, *Los Angeles Times* – 2140 článků, texty z datasetů standardního získávání argumentů 1100 článků

Negative: neutrální texty

- Wikipedia – 3500 článků, věcné zpravodajské zdroje – 3400 článků a další smíšené datasety bez argumentace – 735 článků

Oba: 8800 textů

Hodnocení

Trénovací a testovací část 4 : 1

Table 1. Evaluation results: Nearest Neighbour classification

Method / sources	Precision	Recall	F1
Approach based on keywords	57.2	53.1	55.07
Naive Bayes	59.4	55.0	57.12
Graph-based kNN for reduced CDT (DT only)	65.6	60.4	62.89
Graph-based kNN for reduced CDT (CA only)	62.3	59.5	60.87
Graph-based kNN for full CDT	**83.1**	**75.8**	**79.28**

Table 2. Evaluation results: SVM TK

Method / sources	Precision	Recall	F1
SVM TK for reduced CDT (DT only)	63.6	62.8	63.20
SVM TK for full CDT	82.4	79.61	79.61

Hodnocení

Detekce argumentu vůči zdroji

Table 3. Classification results for individual sources of argumentation (F1 measure)

Method / sources	Newspapers	Textual Complaints (Intense Dataset)	Style and genre recog-nition	Fact and Feeling
Approach based on keywords	52.3	55.2	53.7	54.8
Naive Bayes	57.1	58.3	57.2	59.4
Reduced CDT (DT only)	66.0	63.6	67.9	66.3
Reduced (CA only)	64.5	60.3	62.5	60.9
Full CDT (DT + CA)	77.1	78.8	80.3	79.2

Závěr

Bottleneck jsou prostředky reprezentace.

Vždy je dosaženo mírného zlepšení díky *CDT (DT + CA)*.

Jejich taggovaný soubor zákaznických stížností se dá v budoucnu použít pro další získávání argumentací a klasifikaci textu.

Děkuji za pozornost