OPT: Oslo–Potsdam–Teesside
Pipelining Rules, Rankers, and Classifier Ensembles for Shallow Discourse Parsing

Stephan Oepen[1], Jonathon Read[2], **Tatjana Scheffler**[3], Uladzimir Sidarenka[3,4], Manfred Stede[3], Erik Velldal[1], and Lilja Øvrelid[1]

[1]University of Oslo, Department of Informatics
[2]Teesside University, School of Computing
[3]University of Potsdam, FSP Cognitive Science
[4]Retresco GmbH

tatjana.scheffler@uni-potsdam.de

August 12, 2016

In Short

- good results with classical pipeline
- explicit connectives and arguments: adapted approach from detection of speculation and negation (Velldal et al. 2012, Read et al. 2012)
- cross-validation on training set
- sense disambiguation: ensemble classifier
- $F_1 = 27.77$ on English blind test set

Architecture

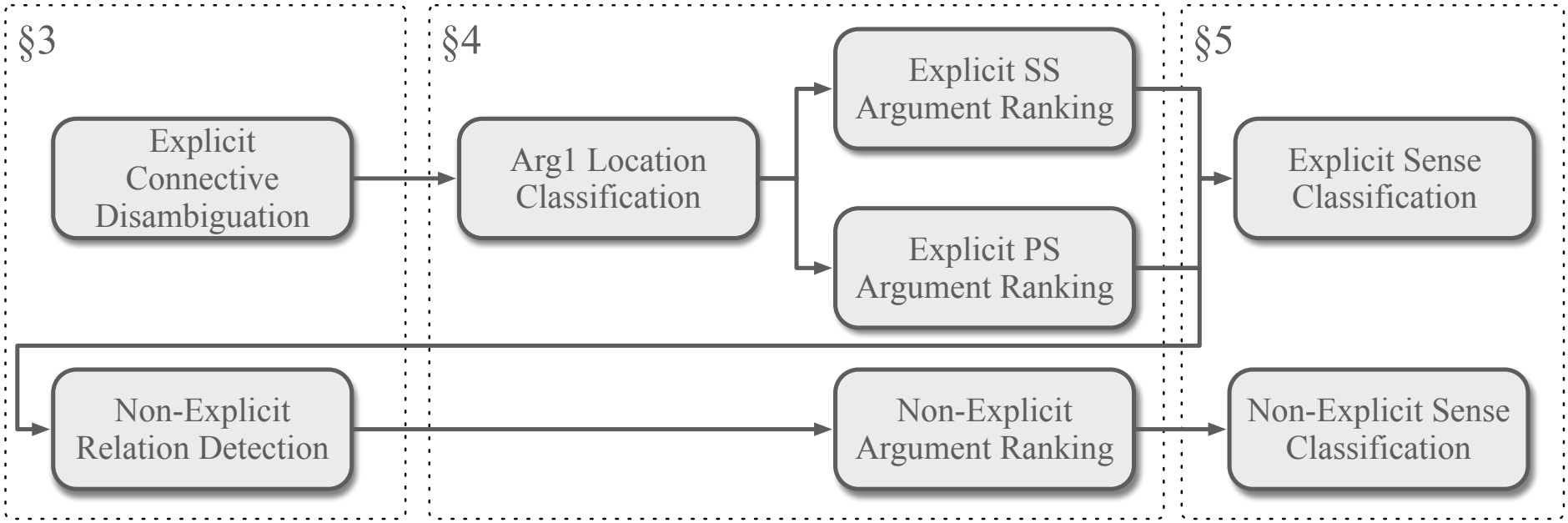

Figure : OPT system overview.

Explicit Connective Detection

- extends the work by Velldal et al. (2012) for identifying expressions of speculation and negation
- disambiguate closed class list of connectives (heads only)
- binary SVM^{light} classifier

Classifier Features

- surface features
  - token and POS n-grams around the candidate (up to ± 5)
- 'parent', 'sibling', 'path', etc. features over PTB-style parse trees
  - from Pitler & Nenkova (2009); Lin et al., (2014); Wang & Lan, (2015)
- feature tuning by ten-fold cross-validation on training set
- final model selection (among some thousand runs):
  - prefer smaller models with less variation across folds
  - test twelve candidate models against development set
- final model:
  - surface features up to 3 tokens before/after candidate
  - full feature conjunction for 'self' and 'parent' categories
  - limited conjunctions for siblings
  - no 'connected context'

Explicit Connective Identification: Results

- $F_1 = 94.4$ on WSJ test set, $F_1 = 91.8$ on blind test set
- comparable to Wang & Lan (2015)
- but well below the best 2016 system (98.9/98.4; Zhongyi Li, Shanghai)

Arguments

- based on work on the scope of speculation and negation (Read et al., 2012)
- assumption: arguments basically correspond to phrases
- approach:
  - extract clausal constituents: S, SBAR, SQ
  - rank them
  - post-editing
- SVM^{light} classifiers; ten-fold cross-validation on training set

Argument Position

Arg2−Arg1	0 (SS)	1 (PS)	2	3	4+
explicit	60.41%	27.98%	5.63%	2.30%	3.66%
non-explicit	2.56%	95.28%	1.64%	0.34%	0.18%

Table : Position of Arg2 relative to Arg1.

Argument Position

Arg2–Arg1	0 (SS)	1 (PS)	2	3	4+
explicit	60.41%	27.98%	5.63%	2.30%	3.66%
non-explicit	2.56%	95.28%	1.64%	0.34%	0.18%

Table : Position of Arg2 relative to Arg1.

- non-explicit relations: Arg1 is in previous sentence (PS) from Arg2
- explicit relations: classifier for PS or same sentence (SS)
  - connective form
  - path from connective to root
  - connective position in sentence (tertiles)
  - POS bigram of connective and following token

Argument Candidate Ranking

- ordinal ranking of clausal constituents
- iteratively build a pool of feature types

	Exp. PS		Exp. SS		Non-Exp.	
	Arg1	**Arg2**	**Arg1**	**Arg2**	**Arg1**	**Arg2**
Connective Form			•			
Connective Category		•				
Connective Precedes				•		
Following Token				•		
Initial Token					•	
Path to Root		•	•		•	•
Path to Connective		•	•	•		
Path to Initial Token					•	•
Preceding Token		•	•		•	•
Production Rules	•	•			•	•
Size						•

Table : Feature types used to describe candidate constituents for argument

Post-Editing Heuristics

	Explicit		Non-Explicit	
	Arg1	**Arg2**	**Arg1**	**Arg2**
Alignment w/o edits	.483	.535	.870	.900
Alignment with edits	.813	.840	.882	.900

Table : Alignment of constituent yield with arguments (in SS or PS).

- initial alignment of full constituent yield with arguments is low
- post-editing rules
  - add conjunction (CC) preceding constituent (`Arg1`)
  - cut clause headed by connective (`Arg1`, explicit, SS)
  - cut constituent-final CC (`Arg1`)
  - cut constituent-final wh-determiner (`Arg1`)
  - cut constituent-initial CC (`Arg2`, explicit)
  - cut relative clause, i.e. SBAR initiated by WHNP/WHADVP
  - cut connective
  - cut initial and final punctuation

Argument Extraction: Results

	WSJ Test Set			Blind Set		
	Arg1	**Arg2**	**Both**	**Arg1**	**Arg2**	**Both**
Explicit (SS)	.683	.817	.590	.647	.783	.519
Explicit (PS)	.623	.663	.462	.611	.832	.505
Explicit (All)	.572	.753	.474	.586	.782	.473
Non-explicit (All)	.744	.743	.593	.640	.758	.539
Overall	.668	.749	.536	.617	.769	.509

Table : Argument extraction results, no error propagation.

Sense Classification

- separate ensemble classifiers for explicit and non-explicit relations:

1. Majority class
2. Wang & Lan (2015)$_{LSVC}$: `LIBLINEAR` SVM classifier
3. Wang & Lan (2015)$_{XGBoost}$: decision trees with gradient boosting, same features

- final prediction label picked from sum of individual classifier probabilities

Sense Classification: Results

	WSJ Test Set			Blind Set		
System	**Exp**	**Non-Exp**	**All**	**Exp**	**Non-Exp**	**All**
2015	**90.79**	34.45	61.27	76.44	**36.29**	**54.76**
Majority	89.30	21.40	54.02	75.91	30.46	51.39
W&L$_{LSVC}$	89.63	**37.18**	**62.29**	**77.86**	33.05	53.66
W&L$_{XGB}$	89.41	34.12	60.64	76.27	34.42	53.62
OPT	89.95	33.53	60.64	76.81	33.66	53.54
LSTM*	89.90	33.76	60.78	77.63	33.69	53.29
OPT*	90.01	**41.12**	**64.70**	77.06	**37.20**	**55.55**

Table : Isolated results for sense classification (the bottom* model was not part of the submission).

Overall Results

- WSJ “test” set and blind test set
- compared to challenge in 2015 and 2016
- error propagation, automatic parses

	WSJ Test Set			Blind Test Set		
	2015 F_1	2016 F_1	OPT F_1	2015 F_1	2016 F_1	OPT F_1
Expl. Conn.	94.8	**98.9**	94.4	91.9	**98.4**	91.8
Expl. Arg1	50.7	**53.8**	52.0	49.7	52.4	**52.4**
Expl. Arg2	**77.4**	76.7	72.6	74.3	75.2	**75.2**
Expl. Arg1Arg2	45.2	**45.3**	43.9	41.4	44.0	**44.0**
Expl. Sense			39.4			34.5
Non-Ex. Arg1	67.2	69.9	**69.9**	60.9	**66.8**	64.6
Non-Ex. Arg2	68.4	71.5	**71.5**	74.6	**79.1**	76.4
Non-Ex. Arg1Arg2	53.1	53.5	**53.5**	50.4	**58.1**	52.0
Non-Ex. Sense			18.0			21.9
All Arg1Arg2	49.4	**49.6**	48.9	46.4	**50.6**	48.2
Overall Parser	29.7	**30.7**	28.2	24.0	27.8	**27.8**

Table : Per-component breakdown of system performance.

Take-Home Messages

- overall, the end-to-end problem is anything but solved

Take-Home Messages

- overall, the end-to-end problem is anything but solved
- adaptation of constituent ranking good fit for argument identification

Take-Home Messages

- overall, the end-to-end problem is anything but solved
- adaptation of constituent ranking good fit for argument identification
- cross-validation has helped reduce over-fitting to WSJ data

Take-Home Messages

- overall, the end-to-end problem is anything but solved
- adaptation of constituent ranking good fit for argument identification
- cross-validation has helped reduce over-fitting to WSJ data
- classifier ensemble improves sense prediction (post-submission results)

Thank you!

Selected References

Read, J., Velldal, E., Øvrelid, L., and Oepen, S. (2012). UiO1. Constituent-based discriminative ranking for negation resolution. In *Proceedings of the 1st Joint Conference on Lexical and Computational Semantics* (pp. 310–318). Montreal, Canada.

Velldal, E., Øvrelid, L., Read, J., and Oepen, S. (2012). Speculation and negation: Rules, rankers and the role of syntax. *Computational Linguistics*, 38(2), 369 – 410.

Non-Explicit Relation Detection

- non-explicit relation between sentences A and B, iff (PDTB):
  - **(i)** A and B are adjacent,
  - **(ii)** A and B are in the same paragraph,
  - **(iii)** A and B are not linked by an explicit connective, and
  - **(iv)** a coherence relation or an entity-based relation holds between them.

Method:

- traverse sentence bigrams (i), (ii)
- check for explicit connectives with Arg1 in PS (iii)
- NoRel (0.6% in PDTB) and AltLex (1.5%) are currently ignored (iv)

Non-Explicit Relation Detection: Results

- module evaluation on gold standard explicit connectives
- $F_1 = 93.2$ on WSJ test set; $P = 89.9, R = 96.8$