Joint prediction in MST style discourse parsing for argumentation mining

Andreas Peldszus   Manfred Stede

Applied Computational Linguistics, University of Potsdam

EMNLP 20.09.2015

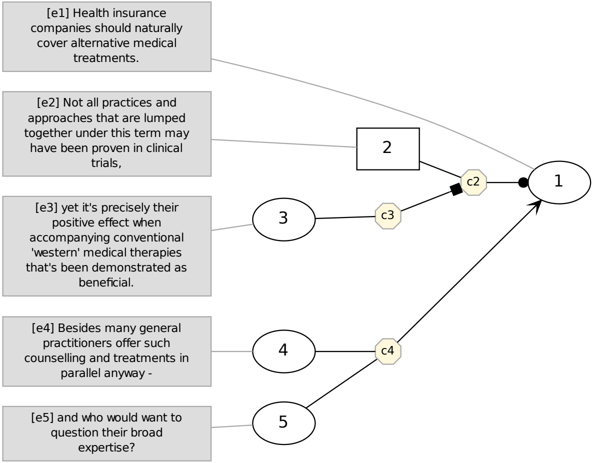
Outline

Outline

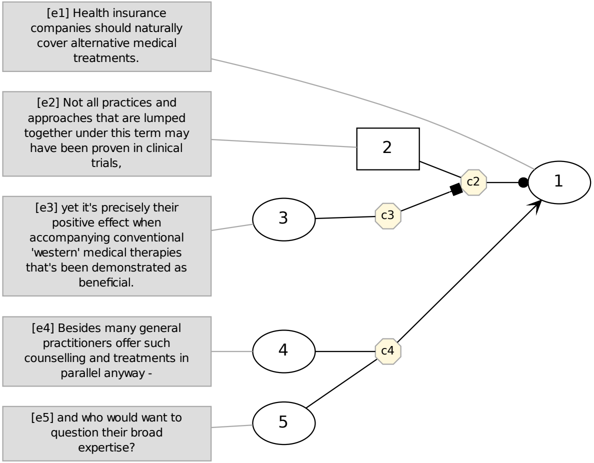
What is argumentation mining?

Health insurance companies should naturally cover alternative medical treatments. Not all practices and approaches that are lumped together under this term may have been proven in clinical trials, yet it's precisely their positive effect when accompanying conventional 'western' medical therapies that's been demonstrated as beneficial. Besides many general practitioners offer such counselling and treatments in parallel anyway - and who would want to question their broad expertise?

What is argumentation mining?



What is argumentation mining?



Tasks:

- EDU segmentation
- ADU segmentation
  resp. argumentative relevance
- ADU type classification
- Relation identification
- Relation type classification

Outline

Dataset: argumentative microtexts

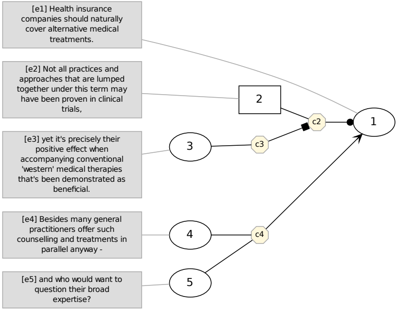
Texts:

- 112 texts: collected in a controlled text generation experiment
- with professional parallel translation to English
- annotated with argumentation structures
- see [Peldszus and Stede, to appear]
- freely available, CC-by-nc-sa license
- `https://github.com/peldszus/arg-microtexts`

Properties:

+ about 5 segments long
+ each segment is arg. relevant
+ explicit main claim
+ at least one possible objection considered

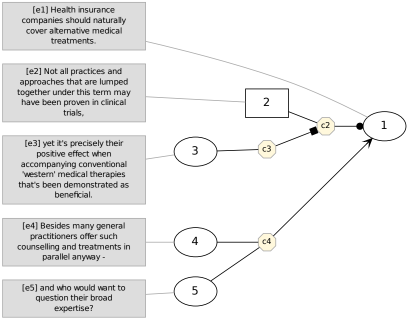
Scheme



Freeman's theory, revised & slightly generalized:
[Freeman, 1991, 2011] [Peldszus and Stede, 2013]

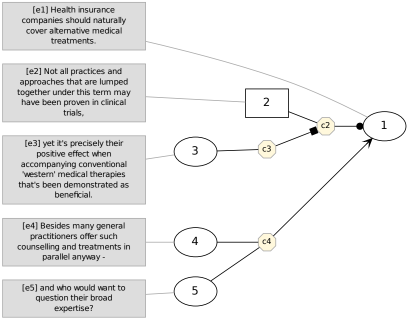
- node types = *argumentative role*
  proponent (presents and defends claims)
  opponent (critically questions)
- link types = *argumentative function*
  support own claims (normally, by example)
  attack other's claims (rebut, undercut)

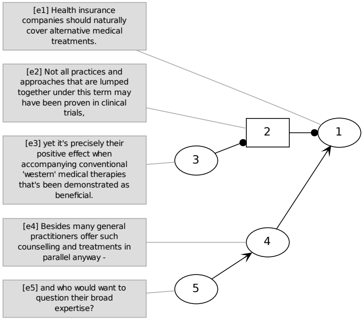
Preprocessing: Graph reduction



From complex structures. . .

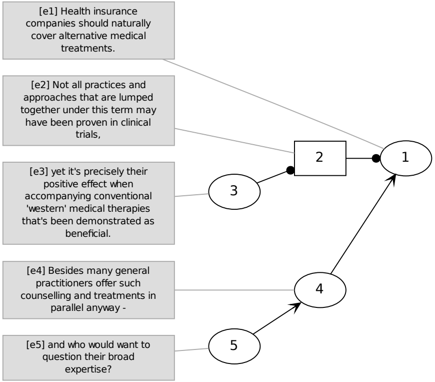
Preprocessing: Graph reduction





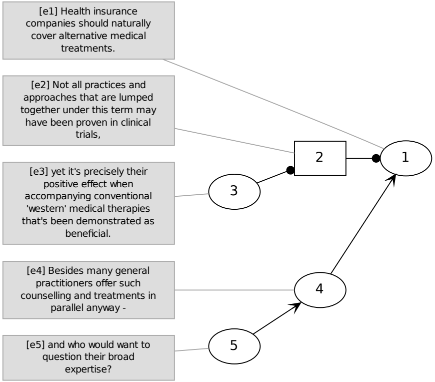
From complex structures... **... to simple support-attack graphs.**

Tasks tackeled in this paper:



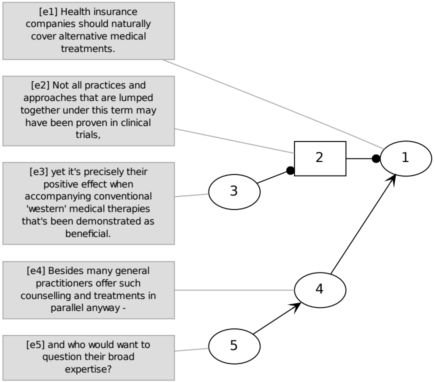
- **attachment (at)**
  464 pairs yes, 2000 pairs no
- **central claim (cc)**
  112 yes, 451 no
- **role (ro)**
  451 proponent, 125 opponent
- **function (fu)**
  290 support, 174 attacks

Tasks tackeled in this paper:



- **attachment (at)**
  464 pairs yes, 2000 pairs no
- **central claim (cc)**
  112 yes, 451 no
- **role (ro)**
  451 proponent, 125 opponent
- **function (fu)**
  290 support, 174 attacks

Tasks tackeled in this paper:



- **attachment (at)**
  464 pairs yes, 2000 pairs no
- **central claim (cc)**
  112 yes, 451 no
- **role (ro)**
  451 proponent, 125 opponent
- **function (fu)**
  290 support, 174 attacks

Outline

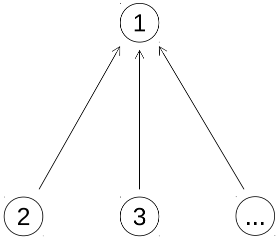
Models: Overview

- Baseline 1
- Baseline 2
- Simple classifier 1
- Simple classifier 2
- Evidence graph 1
- Evidence graph 2
- MSTparser 1
- MSTparser 2
- MSTparser 3
- MSTparser 4
- MSTparser 5

Models: Baseline 1

BL-first: attach to first

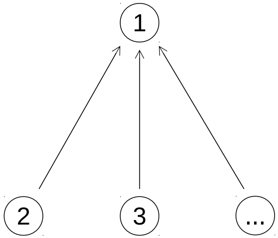
- tendency to state the central claim first in English-speaking debating tradition
- covers convergent argumentation, but not serial argumentation
- in 50 of the 112 texts (44.6%) the first segment is the central claim
- 176 of the 464 relations (37,9%) attach to the first segment



Models: Baseline 1

BL-first: attach to first

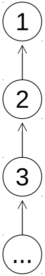
- tendency to state the central claim first in English-speaking debating tradition
- covers convergent argumentation, but not serial argumentation
- in 50 of the 112 texts (44.6%) the first segment is the central claim
- 176 of the 464 relations (37,9%) attach to the first segment



Models: Baseline 2

BL-preced.: attach to preceeding

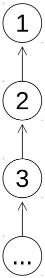
- strong baseline, usually hard to beat
  [Muller et al., 2012]
- covers serial argumentation, but not convergent argumentation
- 210 of all 464 relations (45.3%) attach to the preceeding segment



Models: Baseline 2

BL-preced.: attach to preceeding

- strong baseline, usually hard to beat
  [Muller et al., 2012]
- covers serial argumentation, but not convergent argumentation
- 210 of all 464 relations (45.3%) attach to the preceeding segment



Models: simple classifier

Simple:

- pair-wise classification
- log-loss linear model, SGD training

Procedure:

- predict edge score
- apply classification threshold

Models: simple classifier

Simple:

- pair-wise classification
- log-loss linear model, SGD training

Procedure:

- predict edge score
- apply classification threshold

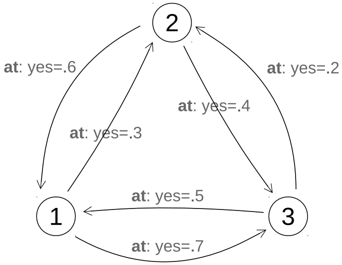
Models: simple classifier

Simple:

- pair-wise classification
- log-loss linear model, SGD training

Procedure:

- predict edge score
- apply classification threshold



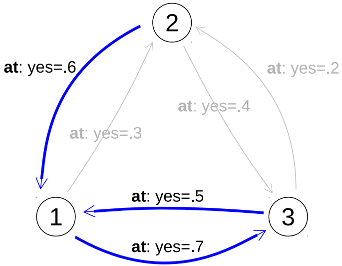
Models: simple classifier

Simple:

- pair-wise classification
- log-loss linear model, SGD training

Procedure:

- predict edge score
- apply classification threshold



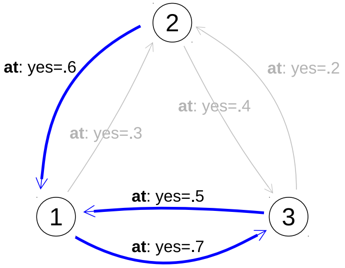
Models: simple classifier

Simple:

- pair-wise classification
- log-loss linear model, SGD training

Procedure:

- predict edge score
- apply classification threshold



Might not be a tree!

Models: features

features from source and target segment:

- lemma unigrams (*) and bigrams
- first three lemma
- POS tags (*)
- main verb mood & tempus
- dependency parse triples
- discourse connectives & relations [Stede, 2002] (*)
- position in text

(*) extracted also from adjacent segment

features of source target pair:

- linerarity (forward, backward)
- absolut distance
- relative distance

Models: features

features from source and target segment:

- lemma unigrams (*) and bigrams
- first three lemma
- POS tags (*)
- main verb mood & tempus
- dependency parse triples
- discourse connectives & relations [Stede, 2002] (*)
- position in text

(*) extracted also from adjacent segment

features of source target pair:

- linerarity (forward, backward)
- absolut distance
- relative distance

Models: features

features from source and target segment:

- lemma unigrams (*) and bigrams
- first three lemma
- POS tags (*)
- main verb mood & tempus
- dependency parse triples
- discourse connectives & relations [Stede, 2002] (*)
- position in text

(*) extracted also from adjacent segment

features of source target pair:

- linerarity (forward, backward)
- absolut distance
- relative distance

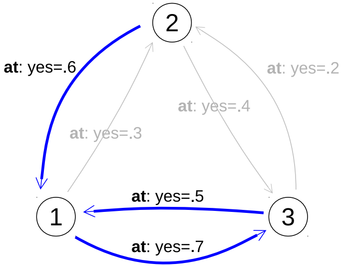
Models: simple classifier

Simple:

- pair-wise classification
- log-loss linear model, SGD training

Procedure:

- predict edge score
- apply classification threshold



***Might* not be a tree!?**

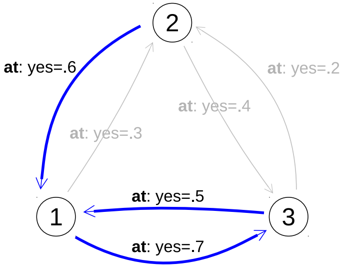
Models: simple classifier

Simple:

- pair-wise classification
- log-loss linear model, SGD training

Procedure:

- predict edge score
- apply classification threshold



Not a tree for 85% of the texts! Too many edges.

Models: simple classifier + MST-decoding

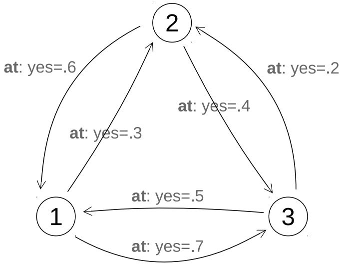
Procedure:

- predict edge score
- apply minimum spanning tree algorithm

Models: simple classifier + MST-decoding

Procedure:

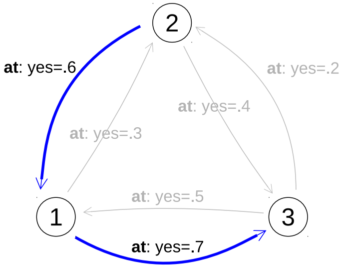
- predict edge score
- apply minimum spanning tree algorithm



Models: simple classifier + MST-decoding

Procedure:

- predict edge score
- apply minimum spanning tree algorithm



Models: simple classifier + MST-decoding

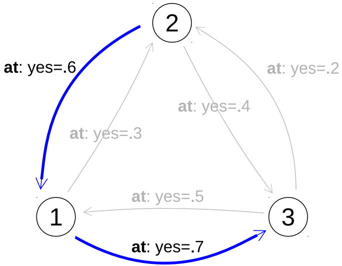
Procedure:

- predict edge score
- apply minimum spanning tree algorithm

  [Chu and Liu, 1965, Edmonds, 1967]

  [McDonald et al., 2005b]

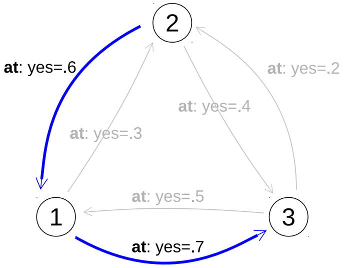
  [Baldridge et al., 2007, Muller et al., 2012]



Models: simple classifier + MST-decoding

Procedure:

- predict edge score
- apply minimum spanning tree algorithm
  [Chu and Liu, 1965, Edmonds, 1967]
  [McDonald et al., 2005b]
  [Baldridge et al., 2007, Muller et al., 2012]



Always a tree!

Attachment classification: results for German

Attachment classification: results for German

	BL-first	BL-preced.	simple	simple+MST
F1 macro	.618	.662	.679	**.688**
attach F1	.380	.452	**.504**	.494
κ	.236	.325	.365	**.377**

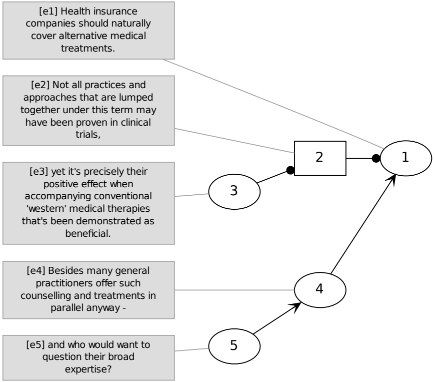
Average scores of 10 repetitions of 5-fold CV.

Attachment classification: results for English

	BL-first	BL-preced.	simple	simple+MST
F1 macro	.618	.662	.663	**.674**
attach F1	.380	.452	**.478**	.470
κ	.236	.325	.333	**.347**

Average scores of 10 repetitions of 5-fold CV.

Why restrict to only one feature of the graph?



Argumentation aspects annotated in the graph:

- **attachment (at)**
- central claim (cc)
- role (ro)
- function (fu)

Models: Joint prediction in one evidence graph

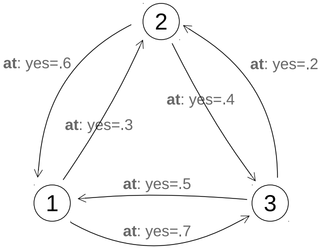
Procedure:

- predict attachment probability
- predict role, function, cc probability
- combine predictions in one score
- apply MST algorithm
- identify central claim
- derive final role class for each segment

Models: Joint prediction in one evidence graph

Procedure:

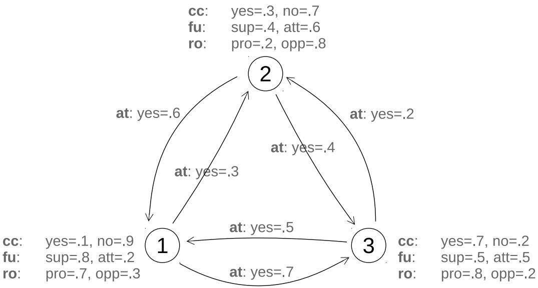
- predict attachment probability
- predict role, function, cc probability
- combine predictions in one score
- apply MST algorithm
- identify central claim
- derive final role class for each segment



Models: Joint prediction in one evidence graph

Procedure:

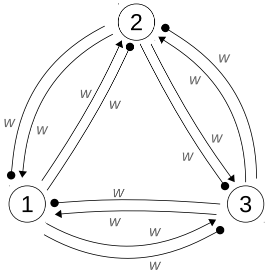
- predict attachment probability
- predict role, function, cc probability
- combine predictions in one score
- apply MST algorithm
- identify central claim
- derive final role class for each segment



Models: Joint prediction in one evidence graph

Procedure:

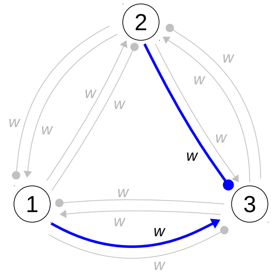
- predict attachment probability
- predict role, function, cc probability
- combine predictions in one score
- apply MST algorithm
- identify central claim
- derive final role class for each segment



Models: Joint prediction in one evidence graph

Procedure:

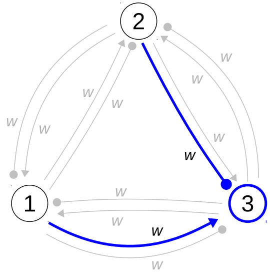
- predict attachment probability
- predict role, function, cc probability
- combine predictions in one score
- apply MST algorithm
- identify central claim
- derive final role class for each segment



Models: Joint prediction in one evidence graph

Procedure:

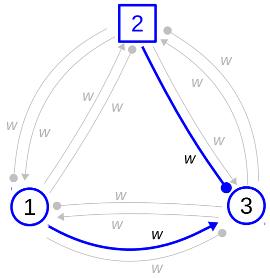
- predict attachment probability
- predict role, function, cc probability
- combine predictions in one score
- apply MST algorithm
- identify central claim
- derive final role class for each segment



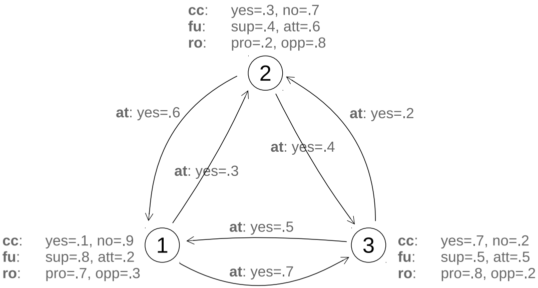
Models: Joint prediction in one evidence graph

Procedure:

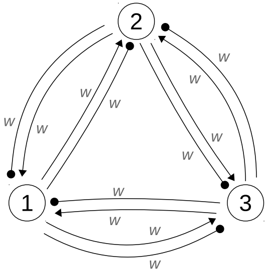
- predict attachment probability
- predict role, function, cc probability
- combine predictions in one score
- apply MST algorithm
- identify central claim
- derive final role class for each segment



Models: Joint prediction in one evidence graph

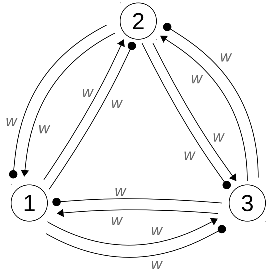


Models: Joint prediction in one evidence graph



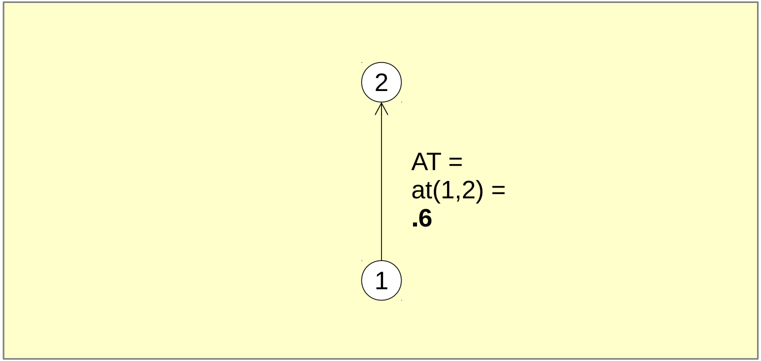
How to move segment-wise predictions into the edge weights?

Models: Joint prediction in one evidence graph



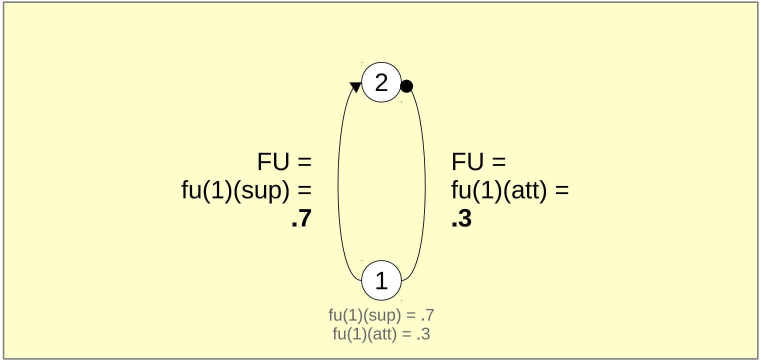
$$w_{i,j} = \frac{\phi_1 \mathrm{RO} + \phi_2 \mathrm{FU} + \phi_3 \mathrm{CC} + \phi_4 \mathrm{AT}}{\sum \phi_n}$$

Models: Joint prediction in one evidence graph



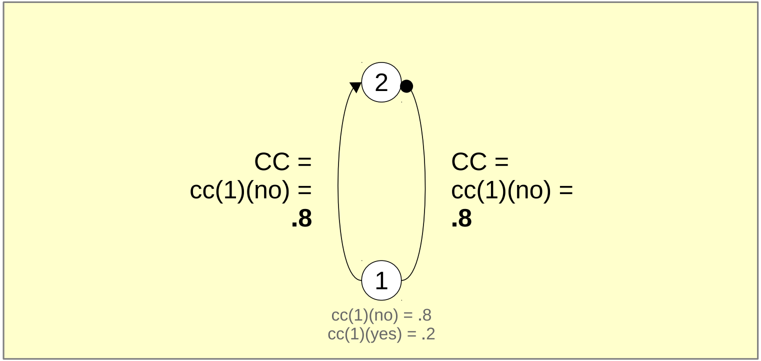
Probability of attachment from source to target

Models: Joint prediction in one evidence graph



Probability of proper function per edge type

Models: Joint prediction in one evidence graph

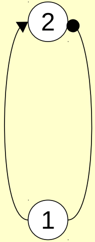


Probability of not being the central claim

Models: Joint prediction in one evidence graph

ro(2)(pro) = .1
ro(2)(opp) = .9

RO =
ro(1)(pro) × ro(2)(pro) +
ro(1)(opp) × ro(2)(opp) =
.9 × .1 + .1 × .9 =
.18

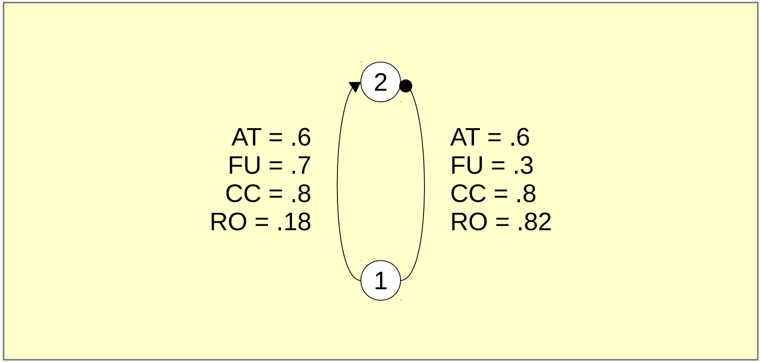


RO =
ro(1)(pro) × ro(2)(opp) +
ro(1)(opp) × ro(2)(pro) =
.9 × .9 + .1 × .1 =
.82

ro(1)(pro) = .9
ro(1)(opp) = .1

Probability of preserved/switched role for sup/att edges

Models: Joint prediction in one evidence graph



$$w_{i,j} = \frac{\phi_1 \mathrm{RO} + \phi_2 \mathrm{FU} + \phi_3 \mathrm{CC} + \phi_4 \mathrm{AT}}{\sum \phi_n}$$

Models: Joint prediction in one evidence graph

EG-equal:

- all base-classifiers weighted equally

EG-best:

- optimize base-classifier weighting with a simple evolutionary search

Models: Joint prediction in one evidence graph

EG-equal:

- all base-classifiers weighted equally

EG-best:

- optimize base-classifier weighting with a simple evolutionary search

Attachment classification: results for German

Attachment classification: results for German

	BL-first	BL-preced.	simple	simple+MST	EG equal	EG best
F1 macro	.618	.662	.679	.688	**.712**	.710
attach F1	.380	.452	.504	.494	**.533**	.530
κ	.236	.325	.365	.377	**.424**	.421

Attachment classification: results for English

	BL-first	BL-preced.	simple	simple+MST	EG equal	EG best
F1 macro	.618	.662	.663	.674	.692	**.693**
attach F1	.380	.452	.478	.470	.501	**.502**
κ	.236	.325	.333	.347	.384	**.386**

Models: Comparison with MSTParser pipelines

Background:

- 1-best MIRA structured learning for non-projective dep. parsing
  [McDonald et al., 2005a]
  [Baldridge et al., 2007]
- same feature sets

Procedure:

- predict edge score
- apply MST decoding
- apply internal or external relation labeller
- derive final role class for each segment

Models: Comparison with MSTParser pipelines

Background:

- 1-best MIRA structured learning for non-projective dep. parsing
  [McDonald et al., 2005a]
  [Baldridge et al., 2007]
- same feature sets

Procedure:

- predict edge score
- apply MST decoding
- apply internal or external relation labeller
- derive final role class for each segment

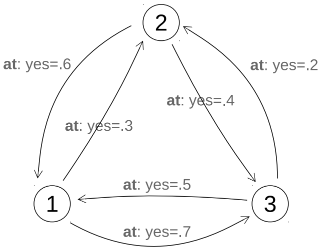
Models: Comparison with MSTParser pipelines

Background:

- 1-best MIRA structured learning for non-projective dep. parsing
  [McDonald et al., 2005a]
  [Baldridge et al., 2007]
- same feature sets

Procedure:

- predict edge score
- apply MST decoding
- apply internal or external relation labeller
- derive final role class for each segment



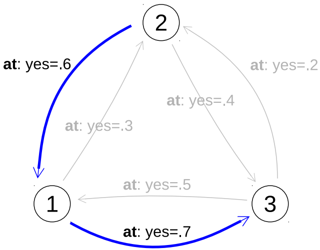
Models: Comparison with MSTParser pipelines

Background:

- 1-best MIRA structured learning for non-projective dep. parsing
  [McDonald et al., 2005a]
  [Baldridge et al., 2007]
- same feature sets

Procedure:

- predict edge score
- apply MST decoding
- apply internal or external relation labeller
- derive final role class for each segment



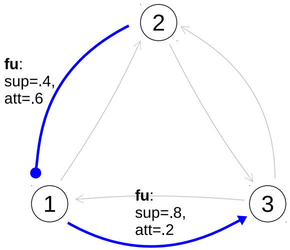
Models: Comparison with MSTParser pipelines

Background:

- 1-best MIRA structured learning for non-projective dep. parsing
  [McDonald et al., 2005a]
  [Baldridge et al., 2007]
- same feature sets

Procedure:

- predict edge score
- apply MST decoding
- apply internal or external relation labeller
- derive final role class for each segment



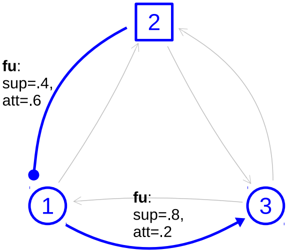
Models: Comparison with MSTParser pipelines

Background:

- 1-best MIRA structured learning for non-projective dep. parsing [McDonald et al., 2005a] [Baldridge et al., 2007]
- same feature sets

Procedure:

- predict edge score
- apply MST decoding
- apply internal or external relation labeller
- derive final role class for each segment



Models: Comparison with MSTParser pipelines

MP:

- normal features
- internal relation labeler

MP+p:

- normal features **plus** base classifier predictions
- internal relation labeler

MP+p+r:

- normal features **plus** base classifier predictions
- **external** relation labeler

Models: Comparison with MSTParser pipelines

MP:

- normal features
- internal relation labeler

MP+p:

- normal features **plus** base classifier predictions
- internal relation labeler

MP+p+r:

- normal features **plus** base classifier predictions
- **external** relation labeler

Models: Comparison with MSTParser pipelines

MP:

- normal features
- internal relation labeler

MP+p:

- normal features **plus** base classifier predictions
- internal relation labeler

MP+p+r:

- normal features **plus** base classifier predictions
- **external** relation labeler

Attachment classification: results for German

Attachment classification: results for German

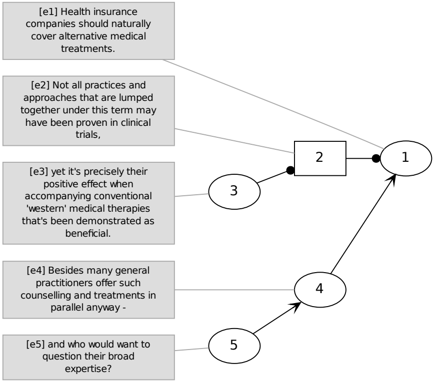
	BL-first	BL-preced.	simple	simple+MST	EG equal	EG best	MP	MP+p
F1 macro	.618	.662	.679	.688	.712	.710	.724	**.728**
attach F1	.380	.452	.504	.494	.533	.530	.553	**.559**
κ	.236	.325	.365	.377	.424	.421	.449	**.456**

Attachment classification: results for English

	BL-first	BL-preced.	simple	simple+MST	EG equal	EG best	MP	MP+p
F1 macro	.618	.662	.663	.674	.692	.693	.707	**.720**
attach F1	.380	.452	.478	.470	.501	.502	.524	**.546**
κ	.236	.325	.333	.347	.384	.386	.414	**.440**

Outline

Tasks tackeled in this paper:



- **attachment (at)**
  464 pairs yes, 2000 pairs no
- **central claim (cc)**
  112 yes, 451 no
- **role (ro)**
  451 proponent, 125 opponent
- **function (fu)**
  290 support, 174 attacks

Results for all levels: German

		simple	EG equal	EG best	MP	MP+p	MP+p+r
at	maF1	.679	.712	.710	.724	**.728**	**.728**
	κ	.365	.424	.421	.449	**.456**	**.456**

Results for all levels: German

		simple	EG equal	EG best	MP	MP+p	MP+p+r
at	maF1	.679	.712	.710	.724	**.728**	**.728**
	κ	.365	.424	.421	.449	**.456**	**.456**
cc	maF1	.849	.879	**.890**	.825	.855	.855
	κ	.698	.759	**.780**	.650	.710	.710

Results for all levels: German

		simple	EG equal	EG best	MP	MP+p	MP+p+r
at	maF1	.679	.712	.710	.724	**.728**	**.728**
	κ	.365	.424	.421	.449	**.456**	**.456**
cc	maF1	.849	.879	**.890**	.825	.855	.855
	κ	.698	.759	**.780**	.650	.710	.710
ro	maF1	**.755**	.737	.734	.464	.477	.669
	κ	**.511**	.477	.472	.014	.022	.340

Results for all levels: German

		simple	EG equal	EG best	MP	MP+p	MP+p+r
at	maF1	.679	.712	.710	.724	**.728**	**.728**
	κ	.365	.424	.421	.449	**.456**	**.456**
cc	maF1	.849	.879	**.890**	.825	.855	.855
	κ	.698	.759	**.780**	.650	.710	.710
ro	maF1	**.755**	.737	.734	.464	.477	.669
	κ	**.511**	.477	.472	.014	.022	.340
fu	maF1	.703	.735	**.736**	.499	.527	.723
	κ	.528	**.573**	.570	.293	.326	.557

Results for all levels: German

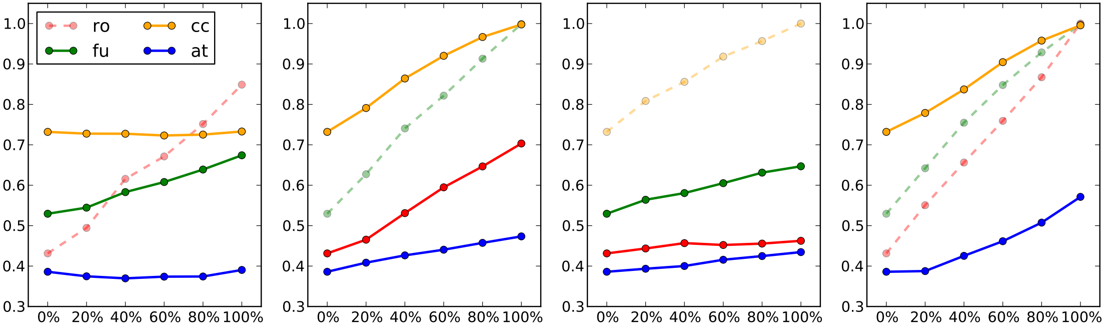
		simple	EG equal	EG best	MP	MP+p	MP+p+r
at	maF1	.679	.712	.710	.724	**.728**	**.728**
	κ	.365	.424	.421	.449	**.456**	**.456**
cc	maF1	.849	.879	**.890**	.825	.855	.855
	κ	.698	.759	**.780**	.650	.710	.710
ro	maF1	**.755**	.737	.734	.464	.477	.669
	κ	**.511**	.477	.472	.014	.022	.340
fu	maF1	.703	.735	**.736**	.499	.527	.723
	κ	.528	**.573**	.570	.293	.326	.557
avg	maF1	.747	.766	**.768**	.628	.647	.744
	κ	.526	.558	**.561**	.352	.379	.516

Results for all levels: English

		simple	EG equal	EG best	MP	MP+p	MP+p+r
at	maF1	.663	.692	.693	.707	**.720**	**.720**
	κ	.333	.384	.386	.414	**.440**	**.440**
cc	maF1	.817	.860	**.869**	.780	.831	.831
	κ	.634	.720	**.737**	.559	.661	.661
ro	maF1	**.750**	.721	.720	.482	.475	.638
	κ	**.502**	.445	.442	.024	.015	.280
fu	maF1	.671	.707	**.710**	.489	.514	.681
	κ	.475	.529	**.530**	.254	.296	.491
avg	maF1	.725	.745	**.748**	.615	.635	.718
	κ	.486	.520	**.524**	.313	.353	.468

Impact of joint prediction

Impact of joint prediction



Simulations of better base classifiers for English (dashed levels artificially improved):
x: number of predictions overwritten with ground truth
y: average κ score in 10 iterations of 5fold CV

Contributions:

- First data-driven model optimizing argumentation structure **globally**.
- First model for argumentation mining **jointly** tackling segment type classification, relation identification and relation type classification.

Contributions:

- First data-driven model optimizing argumentation structure **globally**.
- First model for argumentation mining **jointly** tackling segment type classification, relation identification and relation type classification.

That's it!

- Checkout the corpus:
  `https://github.com/peldszus/arg-microtexts`
- Checkout some evaluations scripts, parameters and (soon) predictions:
  `https://github.com/peldszus/emnlp2015`

Literatur I

Jason Baldridge, Nicholas Asher, and Julie Hunter. Annotation for and robust parsing of discourse structure on unrestricted texts. *Zeitschrift für Sprachwissenschaft*, 26:213–239, 2007.

Y. J. Chu and T. H. Liu. On the shortest arborescence of a directed graph. *Science Sinica*, 14:1396–1400, 1965.

Jack Edmonds. Optimum Branchings. *Journal of Research of the National Bureau of Standards*, 71B:233–240, 1967.

James B. Freeman. *Dialectics and the Macrostructure of Argument*. Foris, Berlin, 1991.

James B. Freeman. *Argument Structure: Representation and Theory*. Argumentation Library (18). Springer, 2011.

Ryan McDonald, Koby Crammer, and Fernando Pereira. Online large-margin training of dependency parsers. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 91–98, Ann Arbor, Michigan, June 2005a. Association for Computational Linguistics.

Ryan McDonald, Fernando Pereira, Kiril Ribarov, and Jan Hajic. Non-projective dependency parsing using spanning tree algorithms. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 523–530, Vancouver, British Columbia, Canada, October 2005b. Association for Computational Linguistics.

Philippe Muller, Stergos Afantenos, Pascal Denis, and Nicholas Asher. Constrained decoding for text-level discourse parsing. In *Proceedings of COLING 2012*, pages 1883–1900, Mumbai, India, December 2012. The COLING 2012 Organizing Committee. URL `http://www.aclweb.org/anthology/C12-1115`.

Andreas Peldszus and Manfred Stede. From argument diagrams to automatic argument mining: A survey. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, 7(1):1–31, 2013.

Andreas Peldszus and Manfred Stede. An annotated corpus of argumentative microtexts. In *Proceedings of the First European Conference on Argumentation: Argumentation and Reasoned Action*, Lisbon, Portugal, June 2015 to appear.

Manfred Stede. DiMLex: A Lexical Approach to Discourse Markers. In Vittorio Di Tomaso Alessandro Lenci, editor, *Exploring the Lexicon - Theory and Computation*. Edizioni dell'Orso, Alessandria, Italy, 2002.