– IN5550 – Neural Methods in Natural Language Processing

Final Exam: Task overview

Stephan Oepen, Lilja Øvrelid,
Vinit Ravishankar & Erik Velldal

University of Oslo

April 25, 2019

Home Exam

General Idea

- Use as guiding metaphor: Preparing a scientific paper for publication.

Home Exam

General Idea

- Use as guiding metaphor: Preparing a scientific paper for publication.

First IN5550 Workshop on Neural NLP (WNNLP 2019)

Home Exam

General Idea

- Use as guiding metaphor: Preparing a scientific paper for publication.

First IN5550 Workshop on Neural NLP (WNNLP 2019)

Standard Process

(1) Experimentation

(2) Analysis

(3) Paper Submission

(4) Reviewing

(5) Camera-Ready Manuscript

(6) Presentation

For Example: The ACL 2019 Conference

WNNLP 2019: Call for Papers and Important Dates

General Constraints

- Four specialized tracks: NLI, NER, Negation Scope, Relation Extraction.
- Long papers: up to nine pages, excluding references, in ACL 2019 style.
- Submitted papers must be anonymous: peer reviewing is double-blind.
- Replicability: Submission backed by code repository (area chairs only).

Schedule

By May 1	Declare team composition and choice of track
May 2	Receive additional, track-specific instructions
May 9	Individual mentoring sessions with Area Chairs
May 16	(Strict) Submission deadline for scientific papers
May 17–23	Reviewing period: Each student reviews two papers
May 27	Area Chairs make and announce acceptance decisions
June 2	Camera-ready manuscripts due, with requested revisions
June 13	Short oral presentations at the workshop

Requirements

- Empirial/experimental
  - some systematic exploration of relevant parameter space, e.g. motivate choice of hyperparameters
  - comparison to reasonable baseline/previous work; explain choice of baseline or points of comparison
- Replicable: everything relevant to re-produce in Microsoft GitHub
- Analytical/reflective
  - relate to previous work
  - meaningful discussion of results
  - 'negative' results can be interesting too
  - discuss some examples: look at the data
  - error analysis

WNNLP 2019: Programme Committee

General Chair

- Andrey Kutuzov

Area Chairs

- Natural Language Inference: Vinit Ravishankar
- Named Entity Recognition: Erik Velldal
- Negation Scope: Stephan Oepen
- Relation Extraction: Lilja Øvrelid & Farhad Nooralahzadeh

Peer Reviewers

- All students who have submitted a scientific paper

Track 1: Named Entity Recognition

- NER: The task of identifying and categorizing proper names in text.
- Typical categories: persons, organizations, locations, geo-political entities, products, events, etc.
- Example from NorNE which is the corpus we will be using:

	ORG					GPE_LOC	
Den	internasjonale	domstolen	har	sete	i	Haag	.
The	*International*	*Court of Justice*	*has*	*its seat*	*in*	*The Hague*	*.*

Class labels

- Abstractly a sequence segmentation task,
- but in practice solved as a sequence labeling problem,
- assigning per-word labels according to some variant of the BIO scheme

B-ORG	I-ORG	I-ORG	O	O	O	B-GPE_LOC	O
Den	internasjonale	domstolen	har	sete	i	Haag	.

NorNE

- First publicly available NER dataset for Norwegian; joint effort between LTG, Schibsted and Språkbanken / the National Library.
- Named entity annotations added to NDT.
- A total of ~ 311K tokens, of which ~ 20K form part of a NE.
- Distributed in the CoNLL-U format using the BIO labeling scheme. Simplified version:

1	Den	den	DET	B-ORG
2	internasjonale	internasjonal	ADJ	I-ORG
3	domstolen	domstol	NOUN	I-ORG
4	har	ha	VERB	O
5	sete	sete	NOUN	O
6	i	i	ADP	O
7	Haag	Haag	PROPN	B-GPE_LOC
8	.	$.	PUNCT	O

NorNE entity types

Type	Train	Dev	Test	Total
PER	4033	607	560	5200
ORG	2828	400	283	3511
GPE_LOC	2132	258	257	2647
PROD	671	162	71	904
LOC	613	109	103	825
GPE_ORG	388	55	50	493
DRV	519	77	48	644
EVT	131	9	5	145
MISC	8	0	0	0

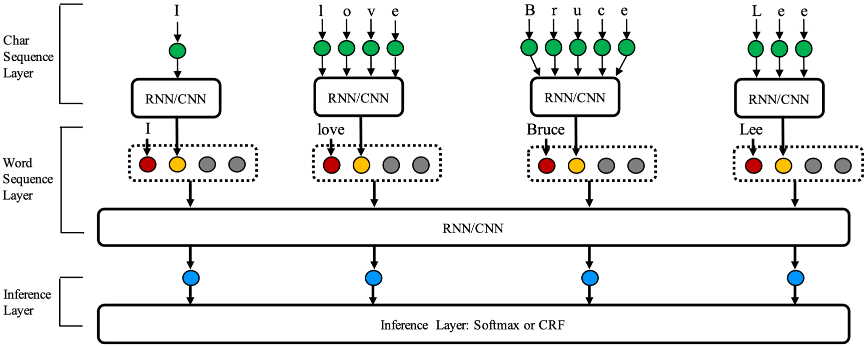
https://github.com/ltgoslo/norne/

Evaluating NER

- https://github.com/davidsbatista/NER-Evaluation
- A common way to evaluate NER is by P, R and F1 at the token-level.
- But evaluating on the entity-level can be more informative.
- Several ways to do this (wording from SemEval 2013 task 9.1 in parens):

- **Exact labeled** ('strict'): The gold annotation and the system output is identical; both the predicted boundary and entity label is correct.
- **Partial labeled** ('type'): Correct label and at least a partial boundary match.
- **Exact unlabeled** ('exact'): Correct boundary, disregarding the label.
- **Partial unlabeled** ('partial'): At least a partial boundary match, disregarding the label.

- Current go-to model for NER: a BiLSTM with a CRF inference layer,
- possibly with a max-pooled character-level CNN feeding into the BiLSTM together with pre-trained word embeddings.



(Image: Jie Yang & Yue Zhang 2018: *NCRF++: An Open-source Neural Sequence Labeling Toolkit*)

Suggested reading on neural seq. modeling

- Jie Yang, Shuailong Liang, & Yue Zhang, 2018
  Design Challenges and Misconceptions in Neural Sequence Labeling
  (Best Paper Award at COLING 2018)
  `https://aclweb.org/anthology/C18-1327`

- Nils Reimers & Iryna Gurevych, 2017
  Optimal Hyperparameters for Deep LSTM-Networks for Sequence Labeling Tasks
  `https://arxiv.org/pdf/1707.06799.pdf`

State-of-the-art leaderboards for NER

- https://nlpprogress.com/english/named_entity_recognition.html
- https://paperswithcode.com/task/named-entity-recognition-ner

Some suggestions to get started with experimentation

- Different label encodings IOB (BIO-1) / BIO-2 / BIOUL (BIOES) etc
- Different label set granularities:
  - 8 entity types in NorNE by default (`MISC` can be ignored)
  - Could be reduced to 7 by collapsing `GPE_LOC` and `GPE_ORG` to `GPE`, or to 6 by mapping them to `LOC` and `ORG`.
- Impact of different parts of the architecture:
  - CRF vs softmax
  - Impact of including a character-level model (e.g. CNN). Tip: isolate evaluation for OOVs.
  - Adding several BiLSTM layers
- Do different evaluation strategies give different relative rankings of different systems?
- Possibilities for transfer / multi-task learning?
- Impact of embedding pre-training (corpus, dim., framework, etc)

Track 2: Natural Language Inference

- How does sentence 2 (hypothesis) relate to sentence 1 (premise)?

Track 2: Natural Language Inference

- How does sentence 2 (hypothesis) relate to sentence 1 (premise)?
- *A man inspects the uniform of a figure in some East Asian country.*

  The man is sleeping

Track 2: Natural Language Inference

- How does sentence 2 (hypothesis) relate to sentence 1 (premise)?
- *A man inspects the uniform of a figure in some East Asian country.*

  The man is sleeping $\rightarrow$ **contradiction**

Track 2: Natural Language Inference

- How does sentence 2 (hypothesis) relate to sentence 1 (premise)?
- *A man inspects the uniform of a figure in some East Asian country.*

  The man is sleeping → **contradiction**
- *A soccer game with multiple males playing.*

  Some men are playing a sport. → **entailment**

Track 2: Natural Language Inference

- How does sentence 2 (hypothesis) relate to sentence 1 (premise)?

Track 2: Natural Language Inference

- How does sentence 2 (hypothesis) relate to sentence 1 (premise)?
- *A man inspects the uniform of a figure in some East Asian country.*

  The man is sleeping

Track 2: Natural Language Inference

- How does sentence 2 (hypothesis) relate to sentence 1 (premise)?
- *A man inspects the uniform of a figure in some East Asian country.*

  The man is sleeping $\rightarrow$ **contradiction**

Track 2: Natural Language Inference

- How does sentence 2 (hypothesis) relate to sentence 1 (premise)?
- *A man inspects the uniform of a figure in some East Asian country.*

  The man is sleeping → **contradiction**
- *A soccer game with multiple males playing.*

  Some men are playing a sport. → **entailment**

Attention

Is attention between the two sentences necessary?

Attention

Is attention between the two sentences necessary?

- "Aye"

  – most people

- "Nay"

  – like two other people

Attention

Is attention between the two sentences necessary?

- "Aye"

  – *most people*

- "Nay"

  – *like two other people*

The ayes mostly have it, but you're going to try both.

Datasets

- **SNLI**: probably the best-known one. Giant leaderboard - https://nlp.stanford.edu/projects/snli/
- **MultiNLI**: Similar to SNLI, but multiple domains. Much harder.
- **BreakingNLI**: the ‘your corpus sucks’ corpus
- **XNLI**: based on MultiNLI, multilingual dev/test portions

Datasets

- **SNLI**: probably the best-known one. Giant leaderboard - https://nlp.stanford.edu/projects/snli/
- **MultiNLI**: Similar to SNLI, but multiple domains. Much harder.
- **BreakingNLI**: the 'your corpus sucks' corpus
- **XNLI**: based on MultiNLI, multilingual dev/test portions
- **NLI5550**: something you can train on a CPU

(Broad) outline

- Two sentences - ‘represent’ them some way, using an encoder
- (optionally) (but not really optionally) use some sort of attention mechanism between them
- Downstream, use a 3-way classifier to guess the label
- Try comparing convolutional encoders to recurrent ones

Compare these approaches - try keeping the number of parameters similar. Describe examples that one system tends to get right better than the other.

Stuff you can look at

- https://arxiv.org/abs/1705.02364 (Conneau et al., 2017) – they learn encoders that they later transfer to other tasks. Interesting encoder design descriptions, you could try one of these out.
- https://www.aclweb.org/anthology/S18-2023 (Poliak et al., 2018) – the authors take the piss out of a lot of existing methods. Great read.
- https://arxiv.org/pdf/1606.01933.pdf (Parikh et al., 2016) – famous attention-y model.
- https://arxiv.org/pdf/1709.04696.pdf (Shen et al., 2017) – slightly more complicated attention-y model. Has a fancy name, therefore probably better.

See also: the granddaddy of all leaderboards – nlpprogress.com/english/natural_language_inference.html

Track 3: Negation Scope

Non-Factuality (and Uncertainty) Very Common in Language

But {this theory would} ⟨not⟩ {work}.

I think, Watson, {a brandy and soda would do him} ⟨no⟩ {harm}.

They were all confederates in {the same} ⟨un⟩{known crime}.

"Found dead ⟨without⟩ {a mark upon him}.

Track 3: Negation Scope

Non-Factuality (and Uncertainty) Very Common in Language

But {this theory would} ⟨not⟩ {work}.

I think, Watson, {a brandy and soda would do him} ⟨no⟩ {harm}.

They were all confederates in {the same} ⟨un⟩{known crime}.

“Found dead ⟨without⟩ {a mark upon him}.

{We have} ⟨never⟩ {gone out ⟨without⟩ {keeping a sharp watch}}, and ⟨no⟩ {one could have escaped our notice}.”

Track 3: Negation Scope

Non-Factuality (and Uncertainty) Very Common in Language

But {this theory would} ⟨not⟩ {work}.

I think, Watson, {a brandy and soda would do him} ⟨no⟩ {harm}.

They were all confederates in {the same} ⟨un⟩{known crime}.

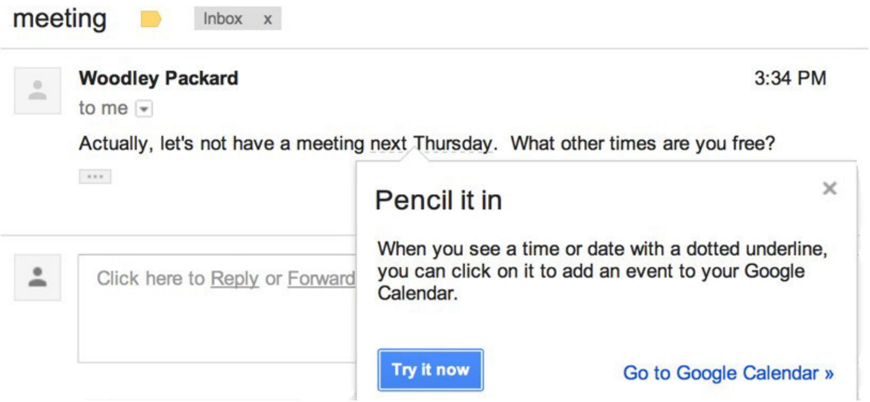
"Found dead ⟨without⟩ {a mark upon him}.

{We have} ⟨never⟩ {gone out ⟨without⟩ {keeping a sharp watch}}, and ⟨no⟩ {one could have escaped our notice}."

Phorbol activation was positively modulated by Ca2+ influx while {TNF alpha activation was} ⟨not⟩.

CoNLL 2010 and *SEM 2012 International Shared Tasks

- Bake-off: Standardized training and test data, evaluation, schedule;
- 20^{+} participants; LTG submissions were top performers in both tasks.


meeting
Inbox x
Woodley Packard
3:34 PM
to me
Actually, let's not have a meeting next Thursday. What other times are you free?
Click here to Reply or Forward
Pencil it in
When you see a time or date with a dotted underline, you can click on it to add an event to your Google Calendar.
Try it now
Go to Google Calendar »

The *SEM 2012 Data (Morante & Daelemans, 2012)

http://www.lrec-conf.org/proceedings/lrec2012/pdf/221_Paper.pdf

ConanDoyle-neg: Annotation of negation in Conan Doyle stories

Roser Morante and Walter Daelemans

CLiPS - University of Antwerp
Prinsstraat 13, B-2000 Antwerp, Belgium
{Roser.Morante,Walter.Daelemans}@ua.ac.be

Abstract

In this paper we present ConanDoyle-neg, a corpus of stories by Conan Doyle annotated with negation information. The negation *cues* and their *scope*, as well as the event or property that is negated have been annotated by two annotators. The inter-annotator agreement is measured in terms of F-scores at scope level. It is higher for cues (94.88 and 92.77), less high for scopes (85.04 and 77.31), and lower for the negated event (79.23 and 80.67). The corpus is publicly available.

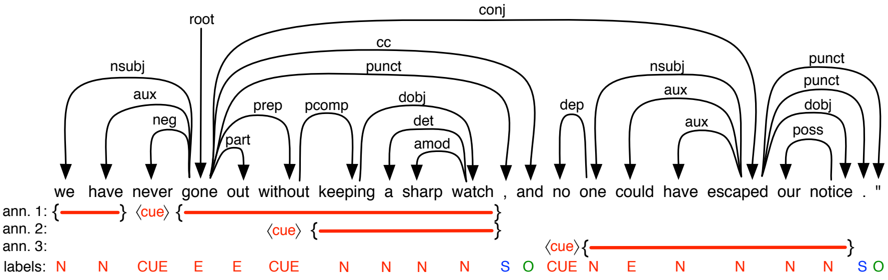
Keywords: Negation, scopes, corpus annotation

1. Introduction

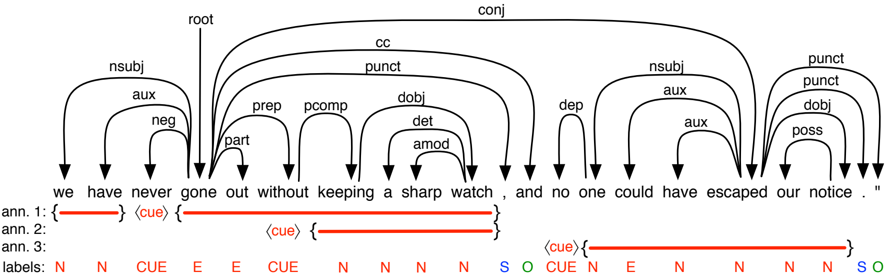
In this paper we present ConanDoyle-neg, a corpus of Conan Doyle stories annotated with negation cues and their scope. The annotated texts are *The Hound of the Baskervilles* (HB) and *The Adventure of Wisteria Lodge* (WL). The original texts are freely available from the Gutenberg Project at `http://www.gutenberg.org/browse/authors/d\#a37238`. The main reason to

nomenon present in all languages. As (Lawler, 2010) puts it, "negation is a linguistic, cognitive, and intellectual phenomenon. Ubiquitous and richly diverse in its manifestations, it is fundamentally important to all human thought". Negation is a frequent phenomenon in language. Tottie reports that negation is twice as frequent in spoken text (27,6 per 1000 words) as in written text (12,8 per 1000 words). Councill et al. (2010) annotate a corpus of product re-

Negation Analysis as a Tagging Task

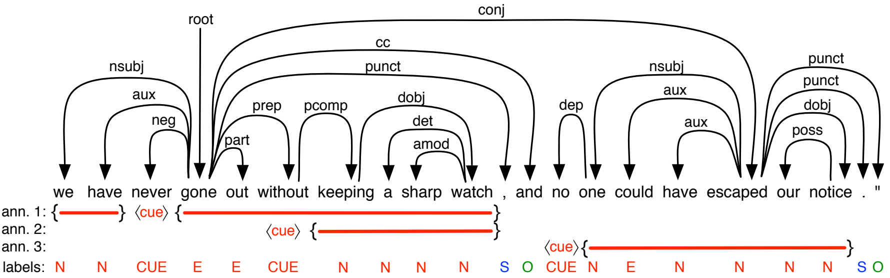


Negation Analysis as a Tagging Task



- Sherlock (Lapponi et al., 2012, 2017) still state of the art today;
- 'flattens out' multiple, potentially overlapping negation instances;
- post-classification: heuristic reconstruction of separate structures.

Negation Analysis as a Tagging Task



- Sherlock (Lapponi et al., 2012, 2017) still state of the art today;
- ‘flattens out’ multiple, potentially overlapping negation instances;
- post-classification: heuristic reconstruction of separate structures.
- To what degree is cue classification a sequence labeling problem?

Probably State of the Art: Lapponi et al. (2017)

http://epe.nlpl.eu/2017/49.pdf

EPE 2017: The Sherlock Negation Resolution Downstream Application

Emanuele Lapponi♣, Stephan Oepen ♣♠, and Lilja Øvrelid♣♠

♣ University of Oslo, Department of Informatics

♠ Center for Advanced Study at the Norwegian Academy of Science and Letters

{emanuel|oe|liljao}@ifi.uio.no

Abstract

This paper describes Sherlock, a generalized update to one of the top-performing systems in the *SEM 2012 shared task on Negation Resolution. The system and the original negation annotations have been adapted to work across different segmentation and morpho-syntactic analysis schemes, making Sherlock suitable to study the downstream effects of different approaches to pre-processing and grammatical analysis on negation resolution.

tion (Björne et al., 2017) and fine-grained opinion analysis (Johansson, 2017), in addition to NR. Although Sherlock and the *SEM 2012 negation data have already been used for extrinsic dependency parsing evaluation, the novelty of the current work lies in the fact that the aforementioned earlier work assumed dependency graphs obtained over uniform, gold-standard sentence and token boundaries, as defined by the original token-level annotations of Morante and Daelemans (2012). In contrast, for use of Sherlock in conjunction with a diverse range of parsers that each start from 'raw', unsegmented https://www.aclweb.org/anthology/P16-1047

Neural Networks For Negation Scope Detection

Federico Fancellu and **Adam Lopez** and **Bonnie Webber**
School of Informatics
University of Edinburgh
11 Crichton Street, Edinburgh
f.fancellu[at]sms.ed.ac.uk, {alopez,bonnie}[at]inf.ed.ac.uk

Abstract

Automatic negation scope detection is a task that has been tackled using different classifiers and heuristics. Most systems are however 1) highly-engineered, 2) English-specific, and 3) only tested on the same genre they were trained on. We start by addressing 1) and 2) using a neural network architecture. Results obtained on data from the *SEM2012 shared task on negation scope detection show that even a simple feed-forward neural network using word-embedding features alone, per-

given the importance of recognizing negation for information extraction from medical records. In more general domains, efforts have been more limited and most of the work centered around the *SEM2012 shared task on automatically detecting negation (§3), despite the recent interest (e.g. machine translation (Wetzel and Bond, 2012; Fancellu and Webber, 2014; Fancellu and Webber, 2015)).

The systems submitted for this shared task, although reaching good overall performance are highly feature-engineered, with some relying on heuristics based on English (Read et al. (2012)) or on tools that are available for a limited number of

Some (Welcome) Simplifications

Separate Sub-Problems in Negation Analysis

- Cue detection   Find negation indicators (sub, single-, or multi-token);
- essentially lexical disambiguation; oftentimes local, binary classification.

Some (Welcome) Simplifications

Separate Sub-Problems in Negation Analysis

- Cue detection   Find negation indicators (sub, single-, or multi-token);
- essentially lexical disambiguation; oftentimes local, binary classification.
- Scope detection   Given one cue, determine sub-strings in its scope;
- structural in principle, but can be approximated as sequence labeling.

Some (Welcome) Simplifications

Separate Sub-Problems in Negation Analysis

- Cue detection Find negation indicators (sub, single-, or multi-token);
- essentially lexical disambiguation; oftentimes local, binary classification.
- Scope detection Given one cue, determine sub-strings in its scope;
- structural in principle, but can be approximated as sequence labeling.
- Event identification within the scope, if factual, find its key 'event'.

Candidate Ways of Dealing with Multiple Negation Instances

- Project onto same sequence of tokens: lose cue–scope correspondence;
- need post-hoc way of reconstructing individual scopes for each cue.

Some (Welcome) Simplifications

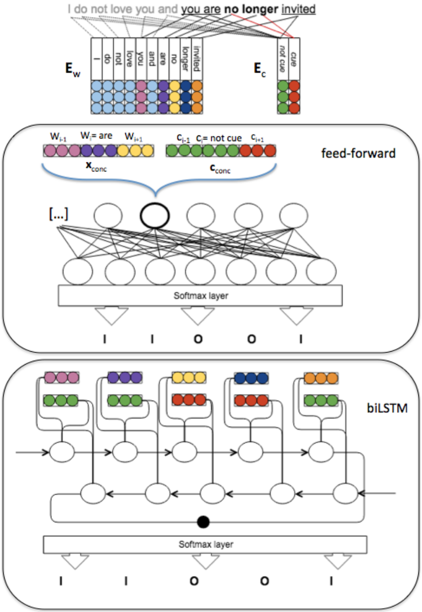
Separate Sub-Problems in Negation Analysis

- Cue detection Find negation indicators (sub, single-, or multi-token);
- essentially lexical disambiguation; oftentimes local, binary classification.
- Scope detection Given one cue, determine sub-strings in its scope;
- structural in principle, but can be approximated as sequence labeling.
- Event identification within the scope, if factual, find its key 'event'.

Candidate Ways of Dealing with Multiple Negation Instances

- Project onto same sequence of tokens: lose cue–scope correspondence;
- need post-hoc way of reconstructing individual scopes for each cue.
- Multiply out: create copy of full sentence for each negation instance;
- risk of presenting 'conflicting evidence', at least for cue detection.

The Architecture of Fancellu et al. (2012)



- Only consider negation scope
- multiplies out multiple instances
- ‘gold’ cue information in input
- Actually, two distinct systems:

(a) independent classification in context of five-grams;

(b) sequence labeling (bi-RNN): binary classification as in-scope

Negation at WNNLP 2019: Our Starting Package

Data and Support Software

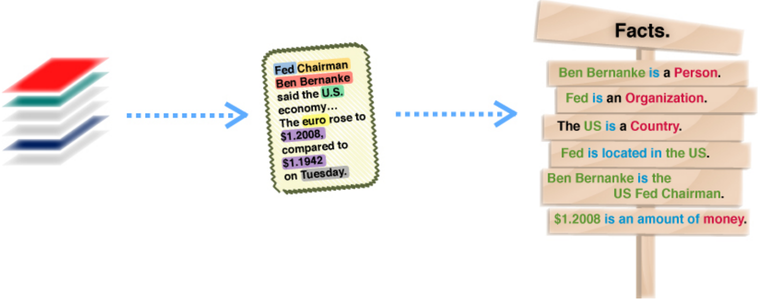
- Four Sherlock Holms stories, carefully annotated with cues and scopes;
- PoS tags and syntactic dependency trees from different parsers;
- easy-to-read JSON serialization; support software to read and write;
- Python interface to standard *SEM 2012 scorer (common metrics).

Possible Research Avenues

- Replicate basic (biLSTM) architecture of Fancellu et al. (2017);
- try out more elaborate labeling schemes (e.g. Lapponi et al., 2017);
- investigate relevance of different PoS tags at different accuracy levels;
- candidate benefits from syntactic structure, e.g. path embeddings;
- actual structured prediction: maximize on whole sequence (e.g. CRF);
- ...

Track 4: Relation extraction

- Identifying relations between entities in text
- Subtask of information extraction pipeline



SemEval 2018 Shared Task

- Semantic relation extraction from scientific texts (Gabor et el., 2018)
- ACL anthology abstracts
- Domain-specific relation set of 6 relations

Usage	*All <u>knowledge sources</u> are treated as <u>feature functions</u>*
Result	*The <u>method</u> yields a <u>performance drop</u> of ...*
Model	*Korean, a <u>verb final language</u> with <u>overt case markers</u>*
Part_Whole	*We use <u>entities</u> extracted from <u>Wikipedia</u>*
Topic	*This <u>paper</u> introduces a new <u>architecture</u>*
Compare	*The correlation of the new <u>measure</u> with <u>human judgment</u> has been investigated ...*

SemEval 2018 Data Set

Relation	Sub-task		Reverse		**Total**
	1.1 & 2	1.2	False	True	
USAGE	483	464	615	332	947
MODEL-FEATURE	326	172	346	152	498
RESULT	72	121	135	58	193
TOPIC	18	240	235	23	258
PART_WHOLE	233	192	273	152	425
COMPARE	95	41	136	-	136
NONE	2315	-	2315	-	2315

Table: Number of instances for each relation in the final dataset

SemEval 2018 Data Set

- We provide an in-house data format
- Pre-processing: XML-parsing, PoS-tagging and dependency parsing
- Each instance contains information about:
  - entity IDs and token spans
  - gold relation and directionality
  - tokenized and lemmatized versions of the sentence
  - PoS-tags and dependency graph
- We also provide domain-specific word embeddings (trained on the ACL anthology)
- Official shared task evaluation script

Data available at
`/projects/nlpl/teaching/uio/in5550/2019/SemEval2018-7`

SemEval 2018 Systems

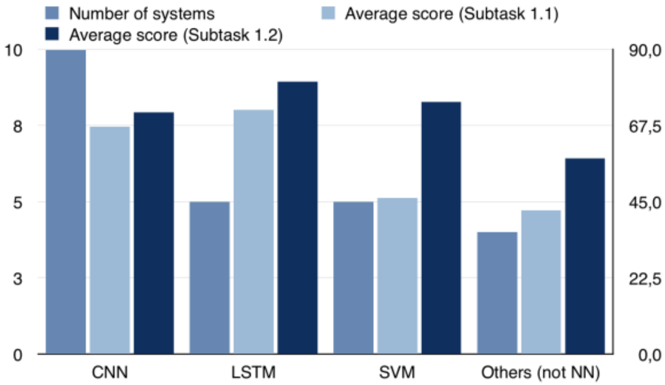


Figure 1: Popularity of methods chosen by participants (as number of systems that used the method, left) and average F1 score obtained for each method (right) in Subtask 1.1 and 1.2.

SemEval 2018 Systems: ETH-DS3Lab

Ensemble system of Rotsztejn et al (2018):

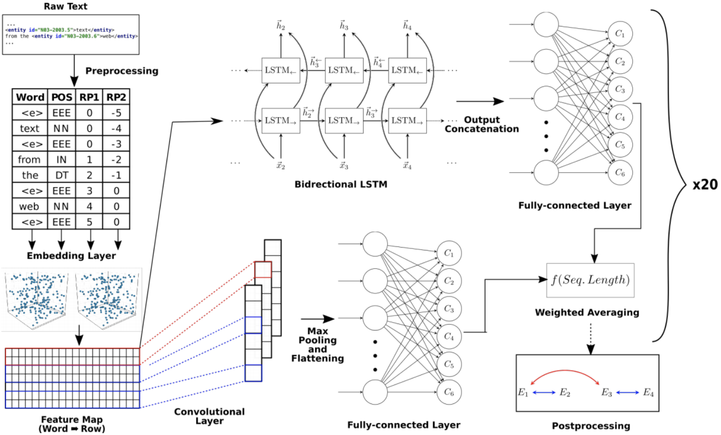


Figure 2: Full pipeline architecture

SemEval 2018 Systems: ETH-DS3Lab

- Combine the strengths of CNNs and RNNs, in addition to a number of other clever tricks
  - domain-specific word embeddings
  - sentence cropping
  - input entity tags
  - PoS-embeddings
  - generate additional data

Suggested reading

- Task website:
  https://competitions.codalab.org/competitions/17422
- Kata Gabor, Davide Buscaldi, Anne-Kathrin Schumann, Behrang QasemiZadeh, Haifa Zaragyouna & Thierry Charnois, 2018
  SemEval-2018 Task 7: Semantic Relation Extraction and Classification in Scientific Papers
  https://aclweb.org/anthology/S18-1111
- Jonathan Rotsztejn, Nora Hollenstein & Ce Zhang, 2018
  ETH-DS3Lab at SemEval-2018 Task7: Effectively Combining Recurrent and Convolutional Neural Networks for Relation Classification and Extraction
  https://aclweb.org/anthology/S18-1112