TEXT MINING

L06. INFORMATION EXTRACTION

1

SUZAN VERBERNE 2022

Universiteit Leiden

ASSIGNMENT 1 – TEXT CLASSIFICATION

- If you didn't make the deadline this week, you can take the resit

	Deadline	Re-sit deadline
Assignment 1	17 October	8 January (maximum grade 6)
Assignment 2	14 November	8 January (maximum grade 6)
Final assignment	8 January	8 February (maximum grade 6)
Written exam	3 January	3 February

- Weight of assignment 1: 10% of total course grade

ASSIGNMENT 1 – TEXT CLASSIFICATION

- We have 9 results tables, that doesn't fit the report
  - You can have 1 table for the three classifiers X the three feature types and experiment with the pre-processing choices for the best classifier.
- Do we need to show results for all methods for all 20 classes?
  - No. If you have room, and you think it is interesting, you could add a table for the individual categories results for the best performing setting.
- How many values of the countvectorizer parameters should we try?
  - You could experiment by manually changing the values and see the effect. E.g. try some extreme values. The goal is that you understand the parameters.
- It takes long to run the experiments
  - It should take minutes (less than 30), not hours.

TODAY'S LECTURE

- Quiz about week 5
- Named Entity Recognition
  - Feature-based models
  - Neural models
- Ontology linking
- Relation extraction

QUIZ ABOUT WEEK 5

- Why should we have multiple human annotators if we create labelled data?
  a. Because we need to estimate the reliability of the data
  b. Because we need to measure the inter-rater agreement between the annotators
  c. Because there is human interpretation involved in the annotation
  d. All of the above

QUIZ ABOUT WEEK 5

- Why should we have multiple human annotators if we create labelled data?
  a. Because we need to estimate the reliability of the data
  b. Because we need to measure the inter-rater agreement between the annotators
  c. Because there is human interpretation involved in the annotation
  d. All of the above

QUIZ ABOUT WEEK 5

- What is the interpretation of Kappa=0?
  a. No agreement
  b. Complete agreement
  c. Measured agreement equal to expected agreement
  d. Undefined

QUIZ ABOUT WEEK 5

- What is the interpretation of Kappa=0?
  a. No agreement
  b. Complete agreement
  c. Measured agreement equal to expected agreement
  d. Undefined

INFORMATION EXTRACTION

INFORMATION EXTRACTION

- IE = central to text mining
- General goal: discover structured information from unstructured or semi-structured text
- Example applications:
  - automatically identify mentions of medications and side effects in electronic health records
  - find person names in bank transactions/electronic health records for the purpose of anonymization
  - find company names, dates and stock market information in economic newspaper texts
  - find scientific references in patents to identify science-market relations

3 EXAMPLES

1. Extracting scientific references from patents

the metabolic products of the cysteine pathway (Dabler et al., Mol. Microbiol., 36, 1101-1112 (2000)). Furthermore, techniques for enhancing

2. Extracting archaeological entities from reports

3 Swifterbant pottery shards from the Middle Neolithic were found.

3. Extracting side effects and coping strategies from patient experiences

Pickle juice reduces my muscle cramps

INFORMATION EXTRACTION TASKS

- Named Entity Recognition (NER)
  (sections 8.3-8.6 in J&M)
- Relation extraction
  (sections 17.1 and 17.2 in J&M)

NAMED ENTITY RECOGNITION

RECOGNIZING ENTITIES

- A named entity is a sequence of words that designates some real-world entity (typically a name), e.g. 'California', 'Steve Jobs' and 'Apple Inc.'
  - General types, occurring in most domains: person, organization, location
  - Extended types (no names): dates, times, monetary values and percentages
  - Domain-specific types, e.g. biomedical entities, archaeological entities

CHALLENGES OF NER

- Ambiguity of segmentation:
  - where are the boundaries of an entity? (e.g. 'King Willem-Alexander of the Netherlands')
- Type ambiguity
  - E.g. The mention 'JFK' can refer to a person, the airport in New York, or any number of schools, bridges, etc.
- Shift of meaning
  - E.g. 'president of the US' refers to Joe Biden, but in a newspaper article from 2011 it refers to Obama and in 2018 to Donald Trump

RECOGNIZING ENTITIES

- We could use a list of names and automatically identify them in the text. Limitations?

RECOGNIZING ENTITIES

- We could use a list of names and automatically identify them in the text. Limitations?
  - Entities are typically multi-word phrases (boundaries?)
  - List is limited (new names, new domains)
  - We would need to add all variants (Trump, Donald Trump, Donald John Trump, President Trump, Mr. Trump, ...)

NAMED ENTITY RECOGNITION (NER)

- NER is a machine learning task based on sequence labelling
  - Word order matters
  - One entity can span multiple words
  - There are multiple ways to refer to the same concept

Mayor Don McLaughlin of Uvalde, Texas, said Tuesday he was frustrated by law enforcement's lack of transparency in the investigation into the mass shooting at Robb Elementary School two weeks ago

- The extracted entities often need to be linked to a standard form (in an ontology or knowledge base)

SEQUENCE LABELLING

- NER is a sequence labelling task
  - sequence = sentence; element = word; label = entity type
  - one label per token
  - the assigned tags capture both the boundary and the type

- Format of training data: IOB tagging
  - Each word gets a label (tag)
  - beginning (B), inside (I) of each entity type
  - and one for tokens outside (O) any entity

Words	IOB Label
American	B-ORG
Airlines	I-ORG
,	O
a	O
unit	O
of	O
AMR	B-ORG
Corp.	I-ORG
,	O
immediately	O
matched	O
the	O
move	O
,	O
spokesman	O
Tim	B-PER
Wagner	I-PER
said	O
.	O

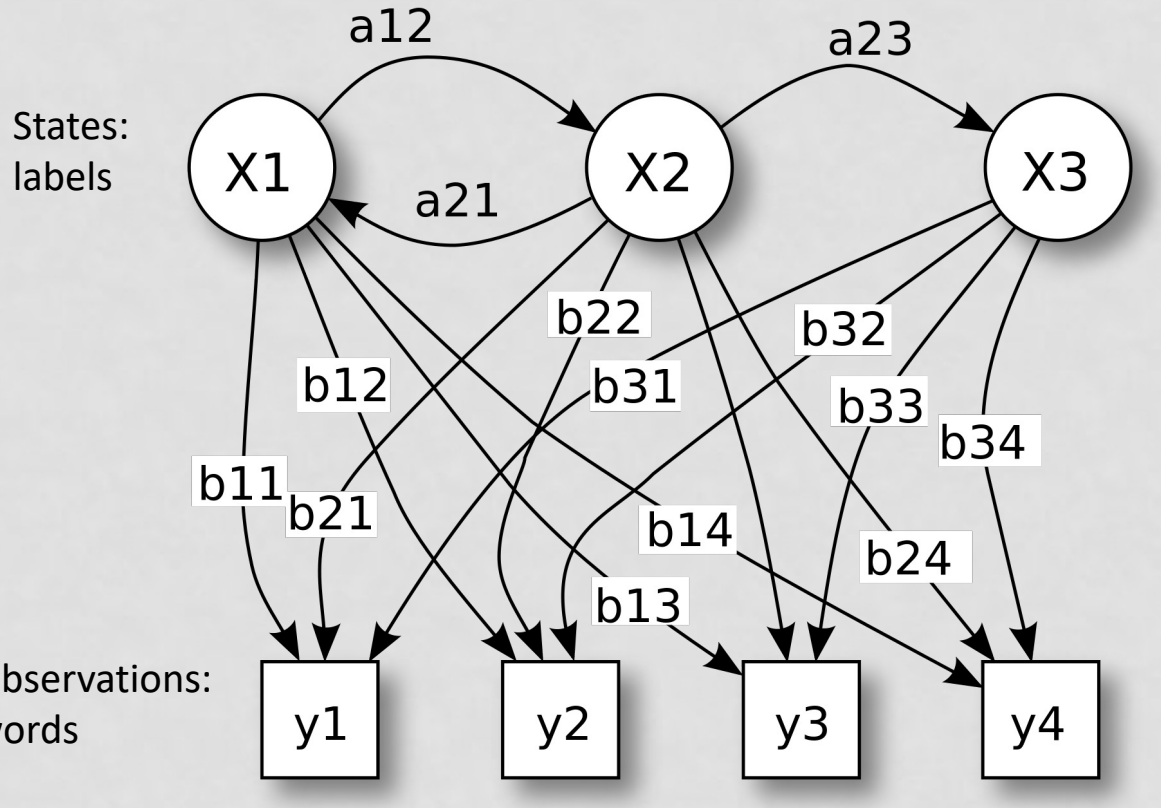
SEQUENCE LABELLING MODELS

J&M CHAPTER 8

HIDDEN MARKOV MODEL (HMM)

- J&M: “The HMM is a classic model that introduces many of the key concepts of sequence modeling that we will see again in more modern models”
- An HMM is a probabilistic sequence model: given a sequence of units (words) it computes a probability distribution over possible sequences of labels and chooses the best label sequence

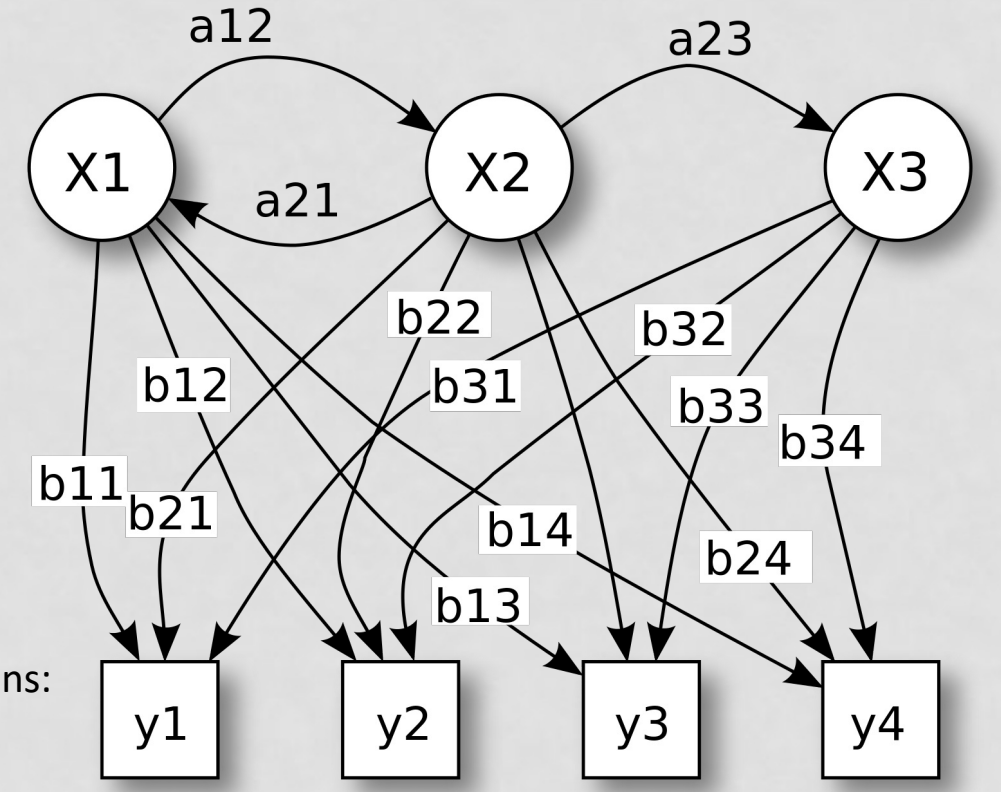
HIDDEN MARKOV MODEL (HMM)



X — states
y — possible observations
a — state transition probabilities
b — output probabilities

J&M section 8.4

HIDDEN MARKOV MODEL (HMM)



X — states
y — possible observations
a — state transition probabilities
b — output probabilities

$$\hat{t}_{1:n} = \underset{t_1 \ldots t_n}{\operatorname{argmax}} P(t_1 \ldots t_n | w_1 \ldots w_n) \approx \underset{t_1 \ldots t_n}{\operatorname{argmax}} \prod_{i=1}^{n} \overbrace{P(w_i|t_i)}^{\text{emission}} \overbrace{P(t_i|t_{i-1})}^{\text{transition}} \quad (8.17)$$

TRAINING HMMS

- In HMM tagging, the probabilities are estimated by counting on a labelled training corpus (remember: Naïve Bayes from lecture 4)
- Task of determining the hidden variables sequence corresponding to the sequence of observations is called decoding

$$\hat{t}_{1:n} = \underset{t_1 \ldots t_n}{\operatorname{argmax}} P(t_1 \ldots t_n | w_1 \ldots w_n) \approx \underset{t_1 \ldots t_n}{\operatorname{argmax}} \prod_{i=1}^{n} \overbrace{P(w_i|t_i)}^{\text{emission}} \overbrace{P(t_i|t_{i-1})}^{\text{transition}} \qquad (8.17)$$

FEATURE-BASED NER

FEATURE-BASED NER

- Supervised learning:
  - Each word represented by a feature vector with information about the word and its context
  - x_i is the word in position i
    - → Create a feature vector for x_i, describing x_i and its context
  - Training data needed: IOB-labeled texts

Mayor	Don	McLaughlin	of	Uvalde	,	Texas	,	said	Tuesday	he
O	B-PER	I-PER	O	B-LOC	I-LOC	I-LOC	O	O	B-TIM	O

FEATURE-BASED NER

- Supervised learning:
  - Each word represented by a feature vector with information about the word and its context
  - x_i is the word in position i
    - → Create a feature vector for x_i, describing x_i and its context

- What features would you use for NER in the general domain (person names, place names, organizations, dates)?

FEATURE-BASED NER

- Commonly used features for sequence labelling NER:

identity of w_i, identity of neighboring words
embeddings for w_i, embeddings for neighboring words
part of speech of w_i, part of speech of neighboring words
presence of w_i in a **gazetteer**
w_i contains a particular prefix (from all prefixes of length ≤ 4)
w_i contains a particular suffix (from all suffixes of length ≤ 4)
word shape of w_i, word shape of neighboring words
short word shape of w_i, short word shape of neighboring words
gazetteer features

Figure 8.15 Typical features for a feature-based NER system.

SIDE STEP: PART-OF-SPEECH TAGGING

- Part-of-speech (POS) = 'category of words that have similar grammatical properties'
  - noun, verb, adjective, adverb
  - pronoun, preposition, conjunction, determiner
  - Example:

    The/DT grand/JJ jury/NN commented/VBD on/IN a/DT number/NN of/IN other/JJ topics/NNS ./.
- Why would the POS of a word be informative for NER?

FEATURE-BASED NER

- Commonly used features for sequence labelling NER:

identity of w_i, identity of neighboring words
embeddings for w_i, embeddings for neighboring words
part of speech of w_i, part of speech of neighboring words
presence of w_i in a **gazetteer**
w_i contains a particular prefix (from all prefixes of length ≤ 4)
w_i contains a particular suffix (from all suffixes of length ≤ 4)
word shape of w_i, word shape of neighboring words
short word shape of w_i, short word shape of neighboring words
gazetteer features

Figure 8.15 Typical features for a feature-based NER system.

FEATURE-BASED NER

- Use of lists:
  - A gazetteer is a list of (place) names
  - Name lists (common first and last person names)
- Word shape features are used to represent the abstract letter pattern of the word by mapping lower-case letters to ‘x’, upper-case to ‘X’, numbers to ’d’, and retaining punctuation. Thus for example I.M.F would map to X.X.X. and DC10-30 would map to XXdd-dd

FEATURE-BASED NER

- Commonly used features for sequence labelling NER:

identity of w_i, identity of neighboring words
embeddings for w_i, embeddings for neighboring words
part of speech of w_i, part of speech of neighboring words
presence of w_i in a **gazetteer**
w_i contains a particular prefix (from all prefixes of length ≤ 4)
w_i contains a particular suffix (from all suffixes of length ≤ 4)
word shape of w_i, word shape of neighboring words
short word shape of w_i, short word shape of neighboring words
gazetteer features

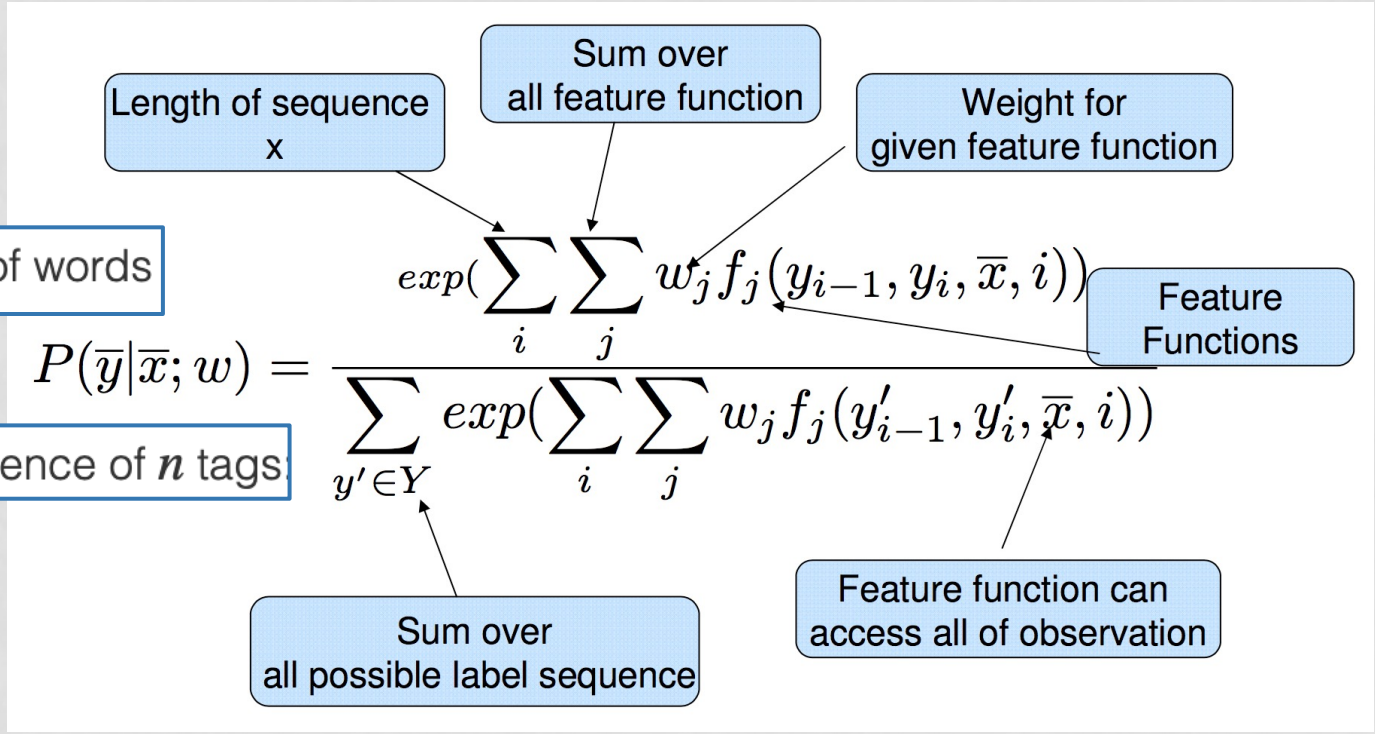
Figure 8.15 Typical features for a feature-based NER system.

- NER task: label all words in the sentence using a sequential model

CONDITIONAL RANDOM FIELDS (CRF)

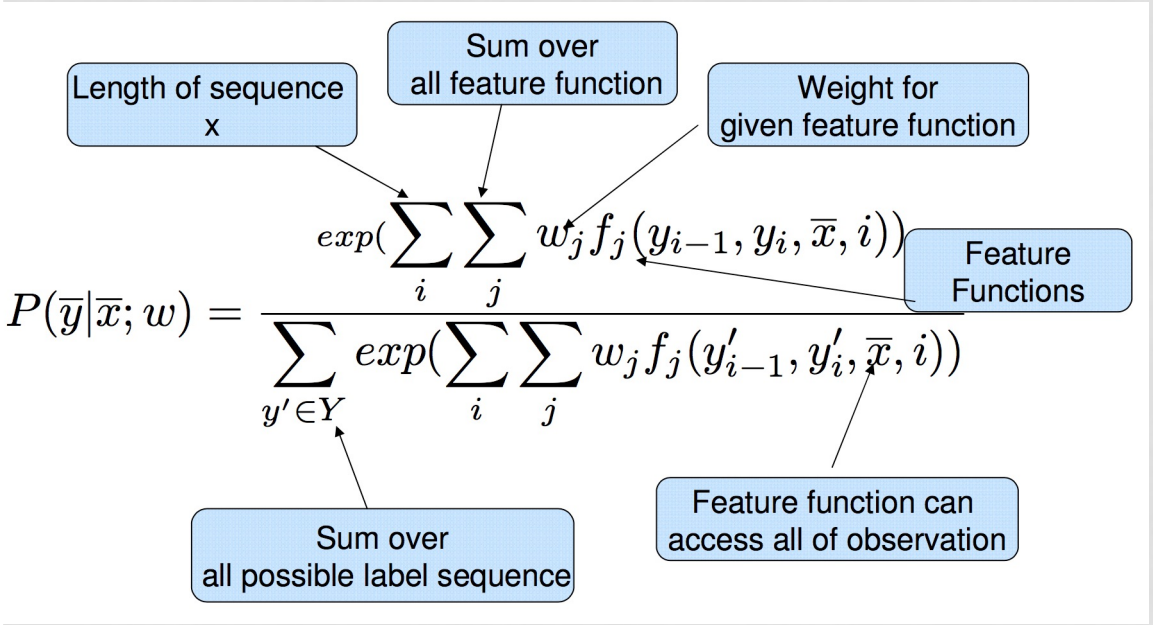
- It is hard for generative models like HMMs to add features directly into the model
- More powerful model: CRF
  - A discriminative undirected probabilistic graphical model
  - Can take rich representations of observations (feature vectors)
  - Takes previous labels and context observations into account
  - Optimizes the sequence as a whole. The probability of the best sequence is computed by the Viterbi algorithm

CONDITIONAL RANDOM FIELDS (CRF)



- http://www.davidsbatista.net/blog/2017/11/13/Conditional_Random_Fields/

CONDITIONAL RANDOM FIELDS (CRF)



Feature functions:

```
features = {
    'bias': 1.0,
    'word.lower()': word.lower(),
    'word[-3:]': word[-3:],
    'word[-2:]': word[-2:],
    'word.isupper()': word.isupper(),
    'word.istitle()': word.istitle(),
    'word.isdigit()': word.isdigit(),
    'postag': postag,
    'postag[:2]': postag[:2],
}
```

- Implementation of CRF in sklearn: https://sklearn-crfsuite.readthedocs.io/en/latest/

NEURAL SEQUENCE MODELS

- Commonly used neural sequence model for NER: biLSTM-CRF
- LSTM = neural architecture with Long Short-Term Memory
- Bi-LSTMs are Recurrent Neural Networks (RNNs)

- RNNs are networks for sequential data.
- The nodes for the current timestamp (token) are connected to the nodes for the previous timestamp

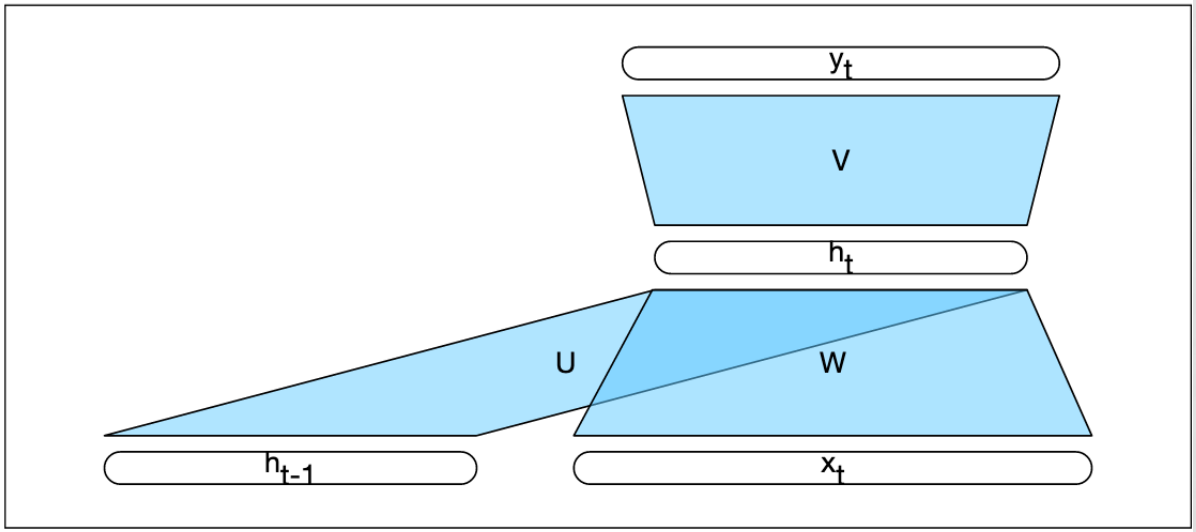


Figure 9.3 Simple recurrent neural network illustrated as a feedforward network.

NEURAL SEQUENCE MODELS

- Commonly used neural sequence model for NER: biLSTM-CRF
- LSTM = neural architecture with Long Short-Term Memory
- Bi-LSTMs are Recurrent Neural Networks (RNNs)
- But for NER the softmax optimization is insufficient:
  - strong constraints for neighboring tokens needed (e.g., the tag I-PER must follow I-PER or B-PER)
  - Solution: Use CRF layer on top of the bi-LSTM output: biLSTM-CRF
- BiLSTM-CRF was the state of the art for NER for some years is still commonly used, in combination with other architectures

TRANSFORMER MODELS

- Current state of the art for Named Entity Recognition: Transformer architectures
- In particular: BERT
- More details next week

STATE OF THE ART FOR NER

- Results on the CONLL-2003 benchmark:

Rank	Model	F1	Extra Training Data	Paper	Code	Result	Year	Tags
1	**ACE + document-context**	94.6	×	Automated Concatenation of Embeddings for Structured Prediction			2021	LSTM, Transformer
2	**Co-regularized LUKE**	94.22	×	Learning from Noisy Labels for Entity-Centric Information Extraction			2021	knowledge distillation
3	**FLERT XLM-R**	94.09	×	FLERT: Document-Level Features for Named Entity Recognition			2020	Transformer
4	**PL-Marker**	94.0	×	Packed Levitated Marker for Entity and Relation Extraction			2021	
5	**LUKE**	93.91	×	LUKE: Deep Contextualized Entity Representations with Entity-aware Self-attention			2020	Transformer

https://paperswithcode.com/sota/named-entity-recognition-ner-on-conll-2003

ONTOLOGY LINKING

NORMALIZATION OF EXTRACTED MENTIONS

- Suppose we have a method to extract:
  - medications and side effects from electronic health records
  - company names and stock market information in newspaper texts
  - scientific references in patents to identify science-market relations

- Multiple extracted mentions can refer to the same concept
- In order to normalize these, we need a list of concepts:
  - Knowledge bases
  - Ontology

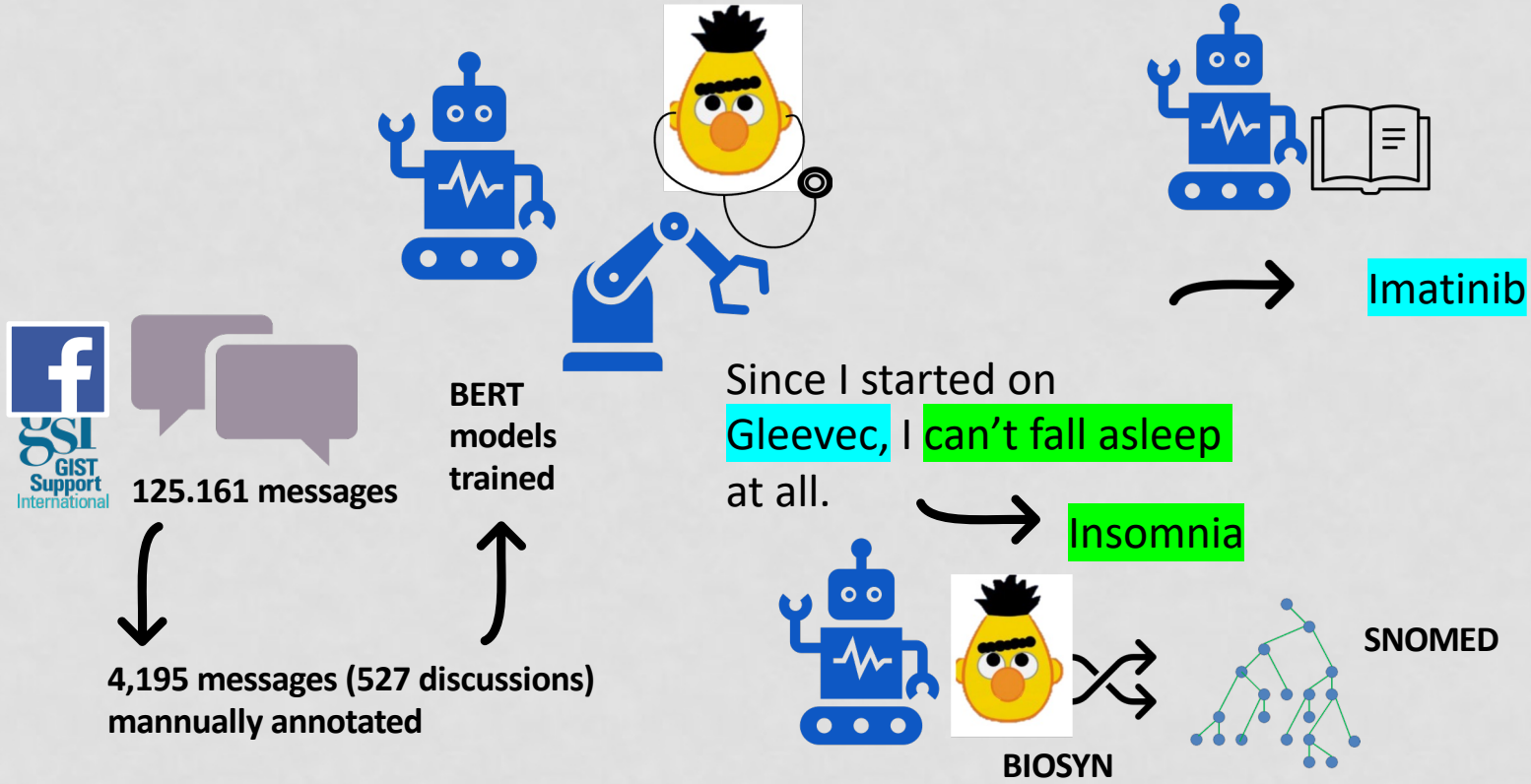
KNOWLEDGE BASES?

- Wikipedia / DBpedia
- IMDB

Domain-specific:

- UMLS/Snomed (medical entities)
- Web of Science (scientific publications)
- … and many other (think of product databases in e-commerce)

EXAMPLE INFORMATION EXTRACTION PIPELINE



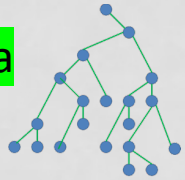
Anne Dirkson, et al. (2022). Automated gathering of real-world data from online patient forums can complement pharmacovigilance for rare cancers. Nature Scientific Reports 12, 10317

ONTOLOGY LINKING APPROACHES

Since I started on Gleevec, I can't fall asleep at all.

Imatinib

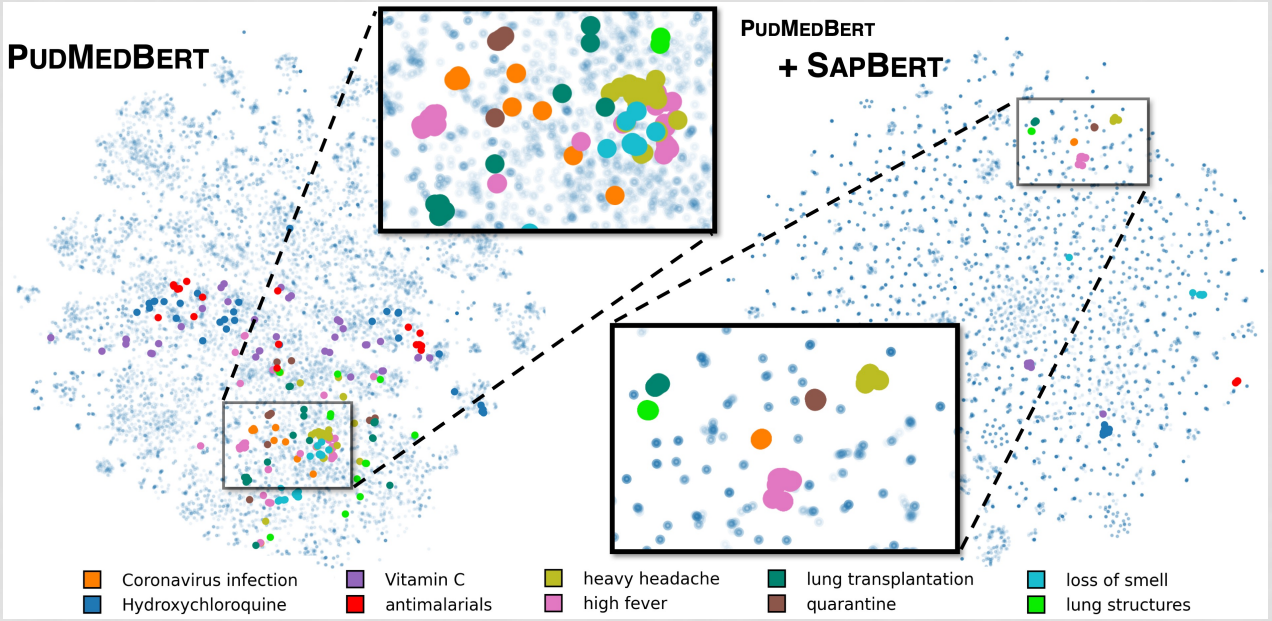
Insomnia



1. Define it as text classification task with the ontology items as labels. Challenges:
   - the label space is huge (Snomed: 311,000 concepts)
   - we don't have training data for all items

ONTOLOGY LINKING APPROACHES

2. Define it as term similarity task:
   - use embeddings trained for synonym detection
   - https://github.com/cambridgeltl/sapbert



Liu et al. (2021). Self-Alignment Pretraining for Biomedical Entity Representations. In *Proceedings of NAACL-HLT*

RELATION EXTRACTION

J&M 17.1 AND 17.2

RELATION EXTRACTION

- Example text with named entities:

Citing high fuel prices, [ORG **United Airlines**] said [TIME **Friday**] it has increased fares by [MONEY **$6**] per round trip on flights to some cities also served by lower-cost carriers. [ORG **American Airlines**], a unit of [ORG **AMR Corp.**], immediately matched the move, spokesman [PER **Tim Wagner**] said. [ORG **United**], a unit of [ORG **UAL Corp.**], said the increase took effect [TIME **Thursday**] and applies to most routes where it competes against discount carriers, such as [LOC **Chicago**] to [LOC **Dallas**] and [LOC **Denver**] to [LOC **San Francisco**].

- Relations:
  - *Tim Wagner* is a spokesman for *American Airlines*
  - *United* is a unit of *UAL Corp.*
  - etc

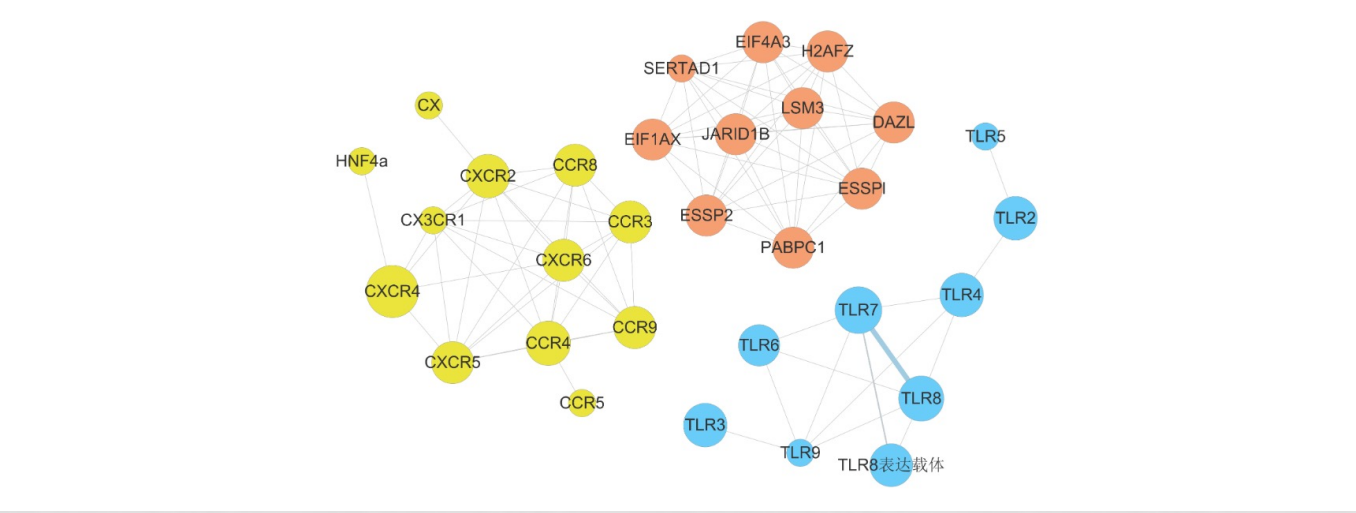
METHODS FOR RELATION EXTRACTION

1. Co-occurrence based
2. Supervised learning (17.2.2)
3. Distant supervision (17.2.4)

- Option 1 relies on the law of big numbers: there will be noise, but the output can still be useful
- Option 2 is the most reliable. However, supervised learning requires labelled data.
- If labelled data is limited, we need option 3.

CO-OCCURRENCE BASED

- Assumption: entities that frequently co-occur are semantically connected
- Use a context window (e.g. sentence) to determine co-occurrence
- We can create a network structure based on this, e.g.:



Yuting Hu and Suzan Verberne (2020). Named Entity Recognition for Chinese biomedical patents. In the Proceedings of the 28th International Conference on Computational Linguistics (COLING), pp 627-637

Universiteit Leiden

SUPERVISED RELATION EXTRACTION

- Assumptions:
  - Two entities, one relation
  - (Relation is verbalized in one sentence, or one passage)
- Relation extraction as classification problem
  1. Find pairs of named entities (usually in the same sentence).
  2. Apply a relation classification on each pair. The classifier can use any supervised technique

SUPERVISED RELATION EXTRACTION

1. E.g. classification with a linear layer on top of an encoder

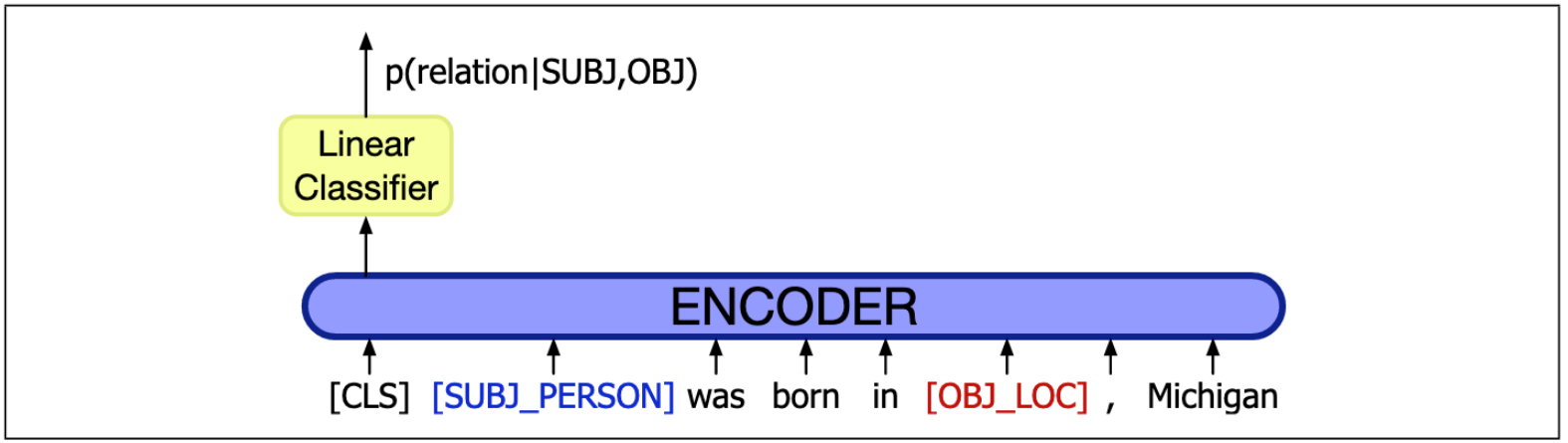


Figure 17.7 Relation extraction as a linear layer on top of an encoder (in this case BERT), with the subject and object entities replaced in the input by their NER tags (Zhang et al. 2017, Joshi et al. 2020).

LIMITED LABELLED DATA

- Suppose we don't have labelled data for relation extraction, but we do have a knowledge base (e.g. IMDB)
- How could you use the knowledge base to identify relations in the text and discover relations that are not yet in the knowledge base?

DISTANT SUPERVISION FOR RELATION EXTRACTION

1. Start with a large, manually created knowledge base (e.g. IMDB)
2. Find occurrences of pairs of related entities from the database in sentences
   - Assumption: If two entities participate in a relation, any sentence that contains these entities express that relation
3. Train a Relation Extraction classifier (supervised) on the found entities and their context
4. Apply the classifier to sentences with yet unconnected other entities in order to find new relations

DISTANT SUPERVISION FOR RELATION EXTRACTION

- The distant supervision paradigm is 13 years old:

[PDF] Distant supervision for relation extraction without labeled data

M Mintz, S Bills, R Snow, D Jurafsky - … of the Joint Conference of the 47th …, 2009 - aclweb.org

Modern models of relation extraction for tasks like ACE are based on supervised learning of relations from small hand-labeled corpora. We investigate an alternative paradigm that does not require labeled corpora, avoiding the domain dependence of ACE-style algorithms, and allowing the use of corpora of any size. Our experiments use Freebase, a large semantic database of several thousand relations, to provide distant supervision. For each pair of entities that appears in some Freebase relation, we find all sentences containing those …

☆ 99 Cited by 2665 Related articles All 27 versions »

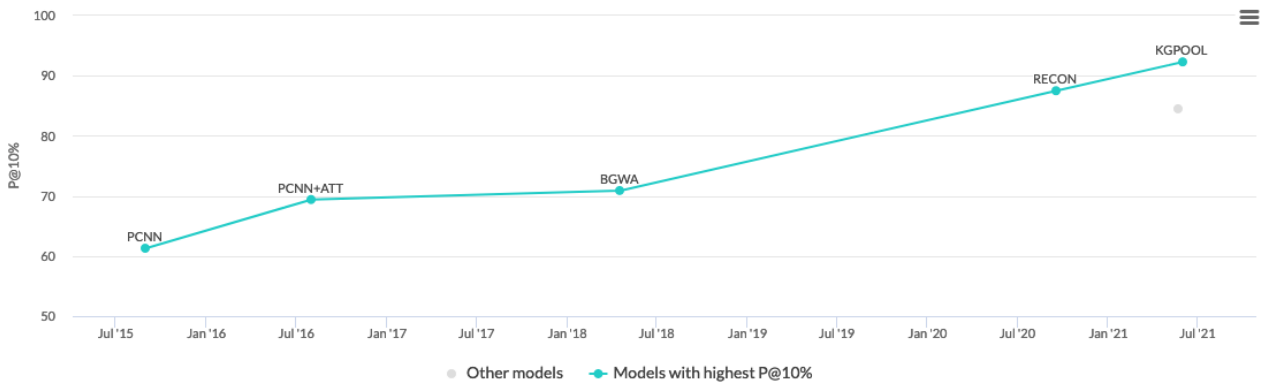
- But still applied in domains with limited labelled data

DISTANT SUPERVISION FOR RELATION EXTRACTION

Relationship Extraction (Distant Supervised) on New York Times Corpus

Leaderboard | Dataset

View P@10% by Date



Filter: untagged

Edit Leaderboard

Rank	Model	P@10%↑	P@30%	AUC	Average Precision	Paper	Code	Result	Year	Tags
1	**KGPOOL**	92.3	86.7			KGPool: Dynamic Knowledge Graph Context Selection for Relation Extraction			2021	

https://paperswithcode.com/sota/relationship-extraction-distant-supervised-on

STATE OF THE ART FOR RELATION EXTRACTION

Rank	Model	RE Micro F1 ↑	RE+ Micro F1	NER Micro F1	Sentence Encoder	Relation classification F1	Cross Sentence	Relation F1	Extra Training Data	Paper	Code	Result	Year
1	**PL-Marker**	73.0	71.1	91.1	ALBERT		Yes		×	Packed Levitated Marker for Entity and Relation Extraction			2021
2	**Ours: cross-sentence ALB**	69.4	67.0	90.9	ALBERT		Yes		×	A Frustratingly Easy Approach for Entity and Relation Extraction			2020
3	**Table-Sequence**	67.6	64.3	89.5	ALBERT		No		×	Two are Better than One: Joint Entity and Relation Extraction with Table-Sequence Encoders			2020

- https://paperswithcode.com/task/relation-extraction

CONCLUSIONS

HOMEWORK

- Read:
  - J&M sections 8.3, 8.4, 8.5, 8.6. Named Entity Recognition
  - J&M sections 17.1 and 17.2. Relation Extraction

HOMEWORK

- Exercise week 6: (after the deadline of assignment 1)

Named entity recognition with CRFsuite

- Follow the tutorial on: https://sklearn-crfsuite.readthedocs.io/en/latest/tutorial.html
- Make sure you understand what the steps mean
- This is **not** a hand-in assignment. You will later complete the 2nd hand-in assignment on this topic (deadline Nov 14)

AFTER THIS LECTURE…

- You can describe the process of Named Entity Recognition (NER) as supervised sequence learning task using 'IOB' labels
- You can list a few commonly used features in NER
- You can explain HMM and CRF for sequence labelling on a conceptual level
- You can explain the task of ontology linking and describe two possible approaches
- You can describe distant supervision for extracting relations between two entities