FAKULTÄT
FÜR !NFORMATIK

Faculty of Informatics

Detecting Multi Word Terms in Patents the same way as Named Entities

Tobias Fink, TU Wien
Linda Andersson, TU Wien and Artificial Researcher IT GmbH
Allan Hanbury, TU Wien

Motivation - Patent Information Retrieval

Technical Terms

blood cell count

battery cell

communication device

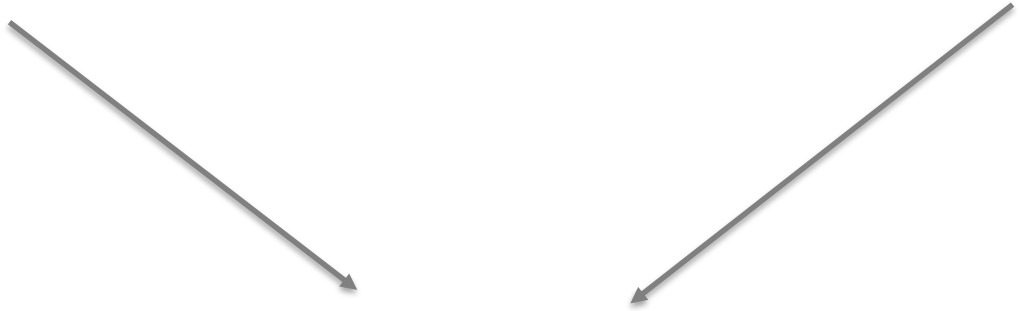
Conventional Methods in the Patent Domain

Pattern:

(**Adj** | **N**)+ **N**

Features:

- Frequencies
- Co-occurrence
- Hand Crafted Features
- …



Machine Learning

Frantzi, Katerina, Sophia Ananiadou, and Hideki Mima. "Automatic recognition of multi-word terms: the c-value/nc-value method." *International journal on digital libraries* 3.2 (2000): 115-130.

Problem Definition - MWTs in Patents

- “Coating” → Verb? Noun?
- “good heat resistance” → MWT boundary?
- MWTs often paraphrased with new MWTs
  → Frequency based statistics unreliable

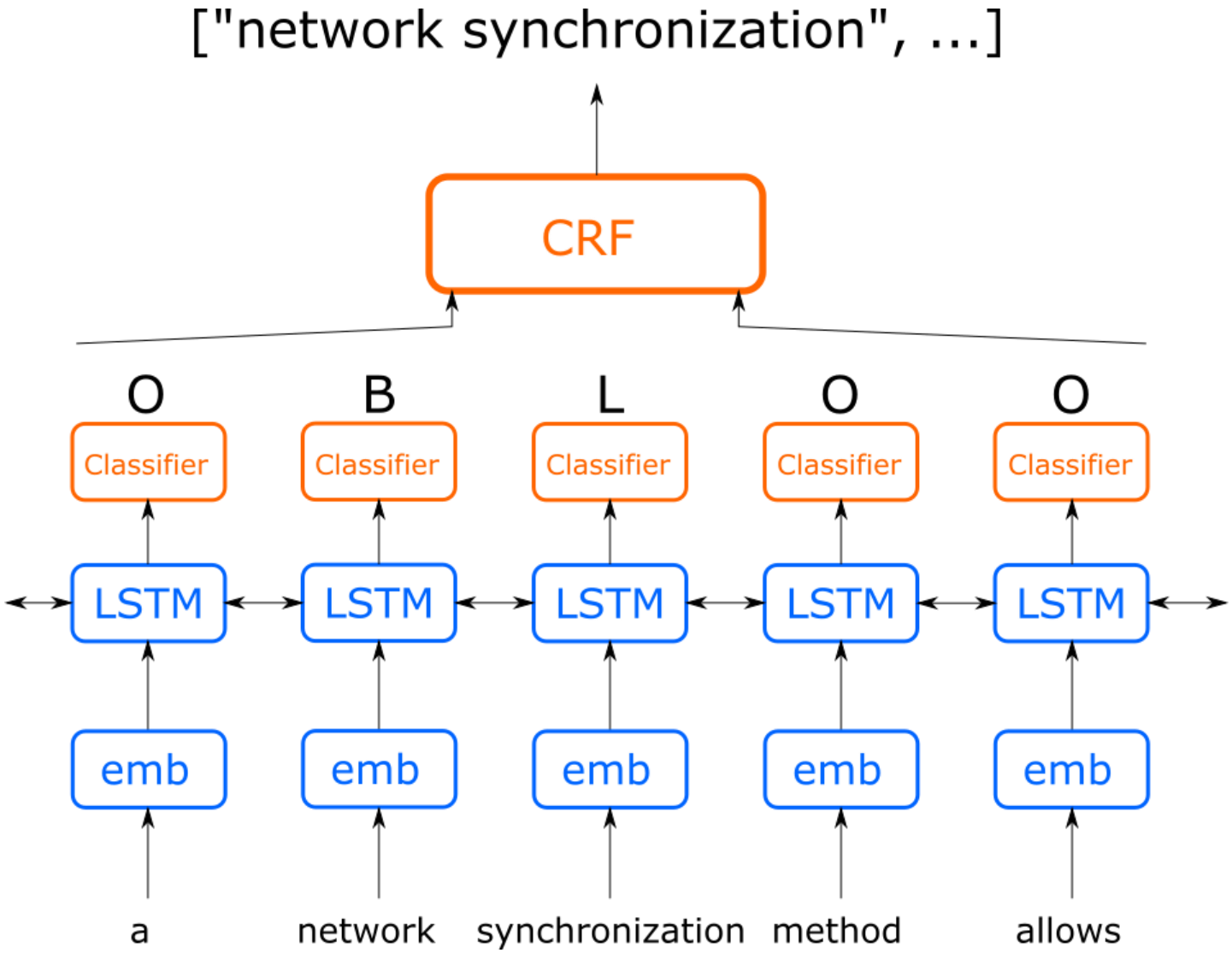
Method Overview

“A **network synchronization** method allows reduced **frequency fluctuations** due to **synchronization control** in a network.“

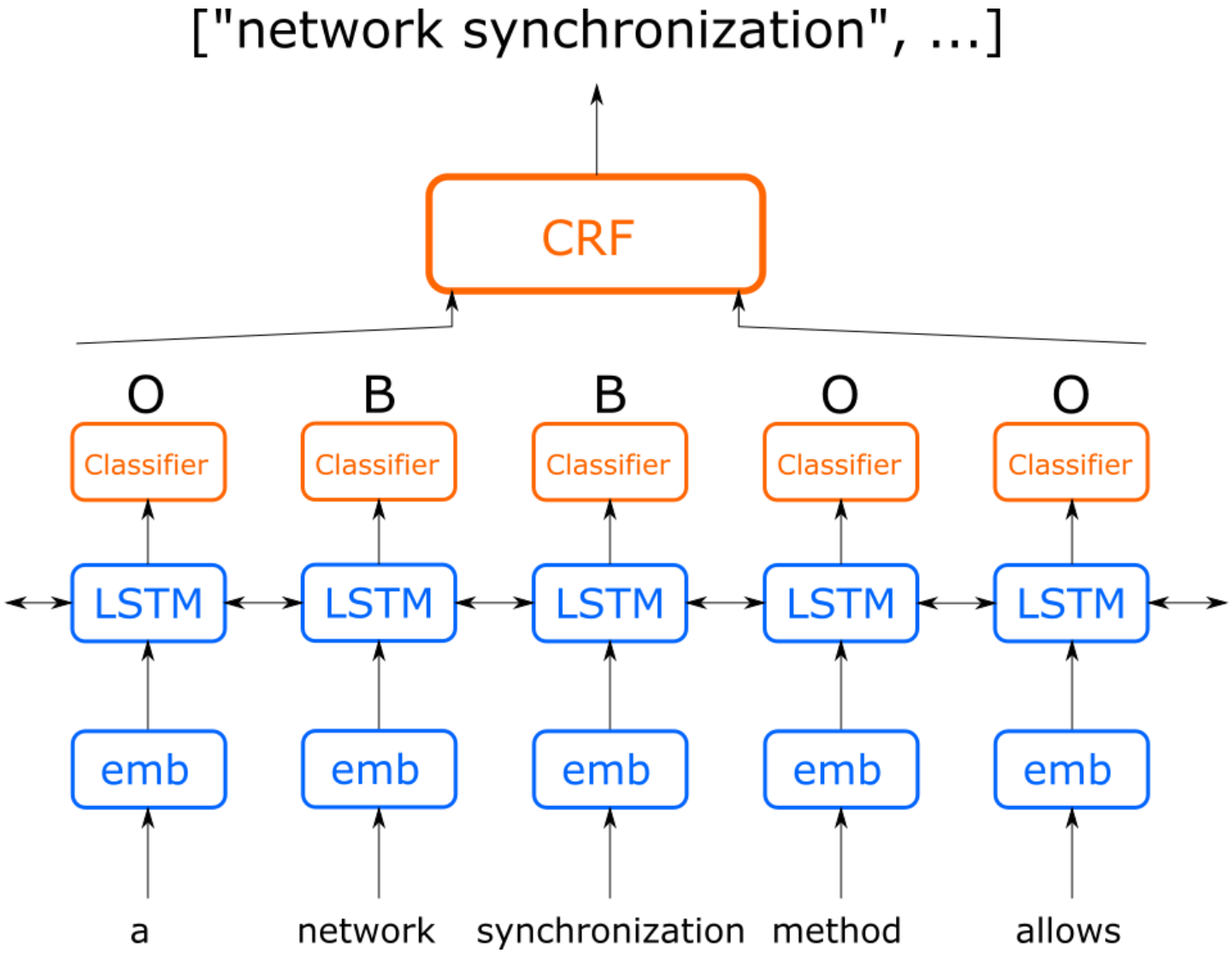
A	network	synchronization	method	allows
		↓		
O	B	L	O	O

Outside Beginning Last

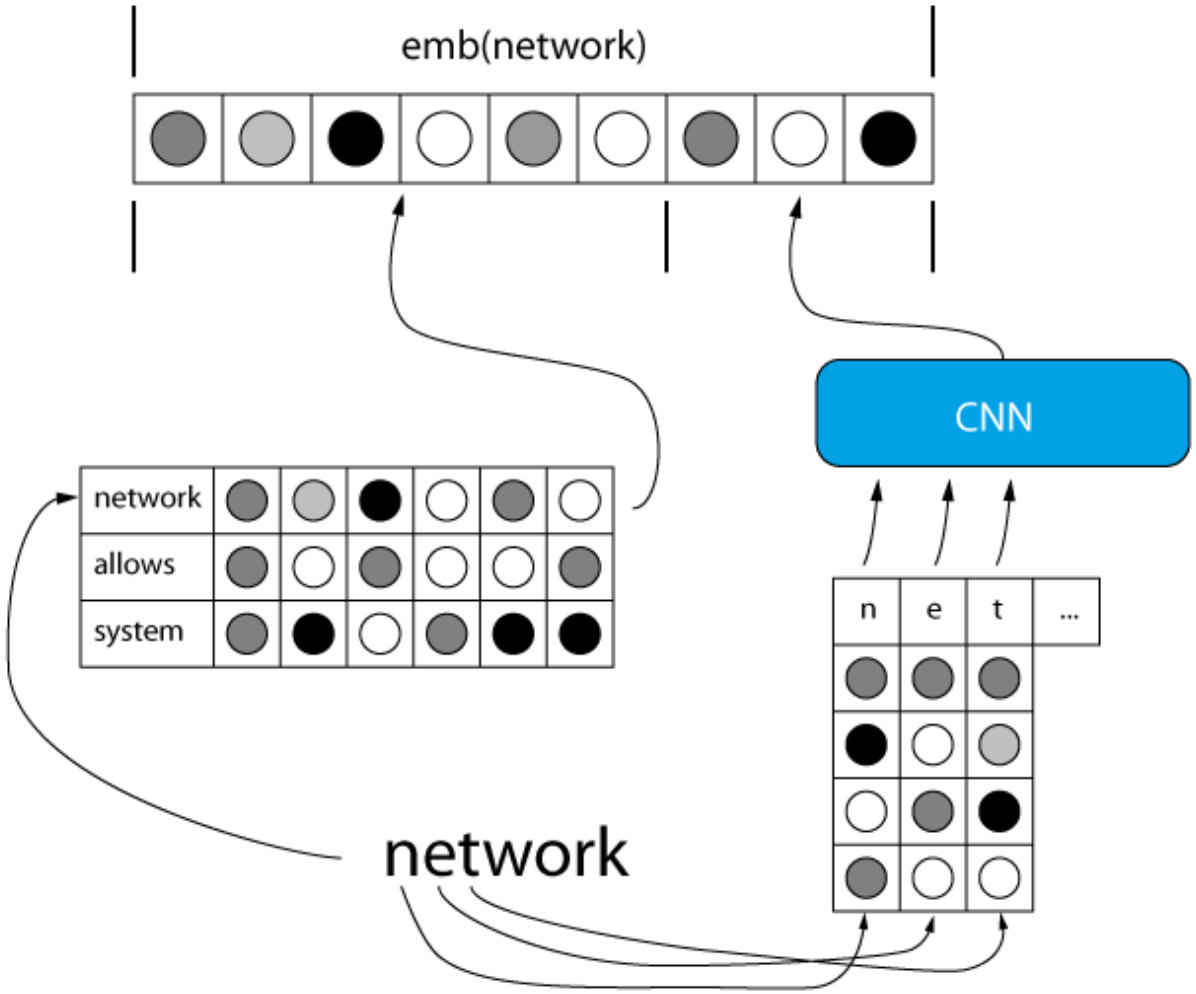
Neural Network Architecture



Neural Network Architecture



Neural Network - Word Embedding



Dataset

- Randomly picked 22 patents with different IPC classes
- Annotated MWTs in the text (single round, 1 annotator)

Number of Tokens	232,065	Avg. MWT dict size (patent)	241
Number of Sentences	10,337	Avg. MWT dict std.dev. (patent)	335
MWT instances	19,465	MWT Occurrence mean	5.12
MWT dictionary size	5,099	MWT Occurrence standard deviation	12.84

Experiments

- Experiments to evaluate impact of:
  - Word Embedding resource (CLEF-IP | Wikipedia)
  - Character based word embedding (CNN component)

- Small Dataset, large variation per patent
  → 11-fold cross validation (20 patents for training,
  2 patents for testing)

- Due to low resources only 2 of 11 folds trained

Results

Model	Avg. Precision	Avg. Recall	Avg. F1
CLEF-IP	**0.75**	**0.74**	**0.74**
Wikipedia	0.71	0.65	0.68
No-CNN	0.67	0.65	0.66
NP-Filter	0.62	0.72	0.66

Out-of-vocabulary Problem

A **distributed system** with a **timing signal path** (22, 90-94, 220) for increased precision in **time synchronization** among **distributed system clocks**.

Detected:
- timing signal path
- time synchronization
- **system clocks**

False Negatives:
- distributed system
- distributed system clocks

Other uses of ‘distributed’: “evenly distributed”, “statistically distributed”

Substring Influences

Method for making **coated metallic cord**

The present invention is intended to create an arrangement for activating and deactivating **automatic noise cancellation** when it is required.

Detected:

- **making coated metallic**
- **deactivating automatic noise cancellation**

False Negatives:

- coated metallic cord
- automatic noise cancellation

“activating” never used, but substring “activ” and “ing” appear in MWTs

Conclusion

- Model learns termhoodness for individual words
- Small dataset leads to out-of-vocabulary problem
- CNN component boosts performance
  → Domain independent connections
- Additional (pre-trained) components might increase performance

Questions?