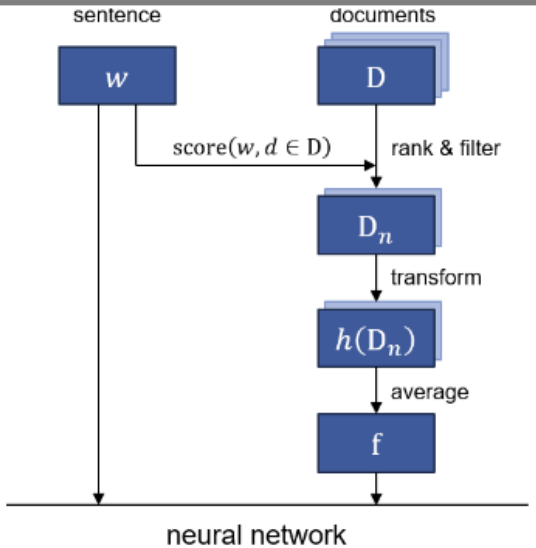
Karlsruhe Institute of Technology

Integrating Encyclopedic Knowledge into Neural Language Models

Yang Zhang, Jan Niehues, Alex Waibel

Interactive Systems Labs
Institute for Anthropomatics and Robotics, Faculty of Informatics



Problem

- Polysemy problem in translation task, e.g.
- "The author does research for his column."
  - "Der Autor betreibt Forschung für seine Säule" →✗
  - "Der Autor betreibt Recherche für seine Kolumne" ✓
- Why does this happen?
  - High translation score
  - Language model fails to capture contextual meaning
- Problem for low-resource languages

Approach

Motivation

- Context can be used to disambiguate
  - author → column/Kolumne
- Can we use other resources to disambiguate?
  - Encyclopedia

Approach

Idea

- Add topic-related, external information into neural language models
  - Long context
  - Add arbitrary information
- Sources
  - Wikipedia
  - zdic.net
- Word level: Wikipedia Categories
- Sentence level: Topic modelling of encyclopaedia documents

➔ Use this to improve rescoring of a SMT system

Contents

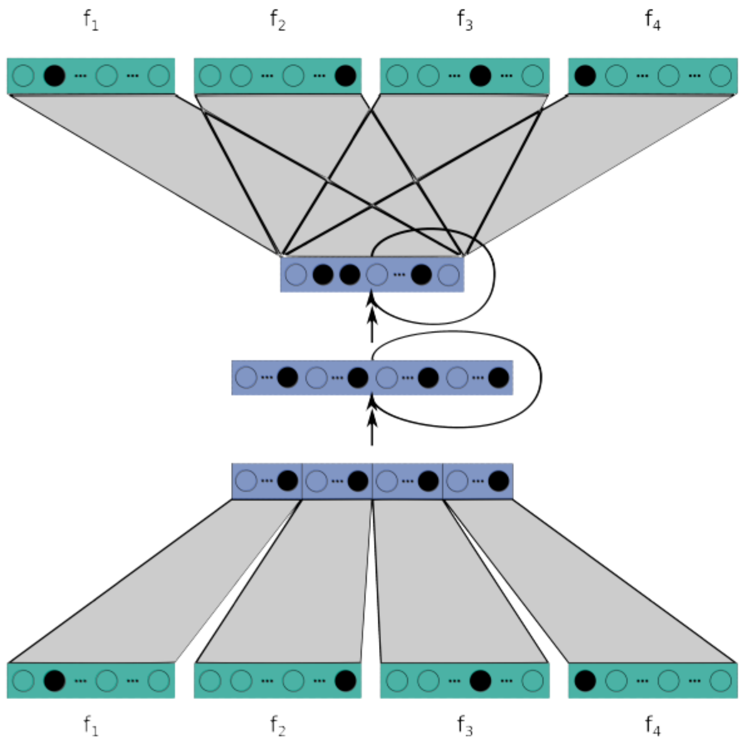
- Motivation
- Integration
- Model
  - Word-level integration
  - Sentence-level integration
- Evaluation
- Conclusion

Rescoring

- Create **n-best lists** for dev/test set
- Retrieve information for encyclopedia
- Score these n-best lists with **new model**
- Add scores to previous n-best scores
- Find model weights (MERT, ListNet)

Neural Network Language Model

- LSTM-based LM
  - 2 layers
- Factored word representation
  - Input
  - Output

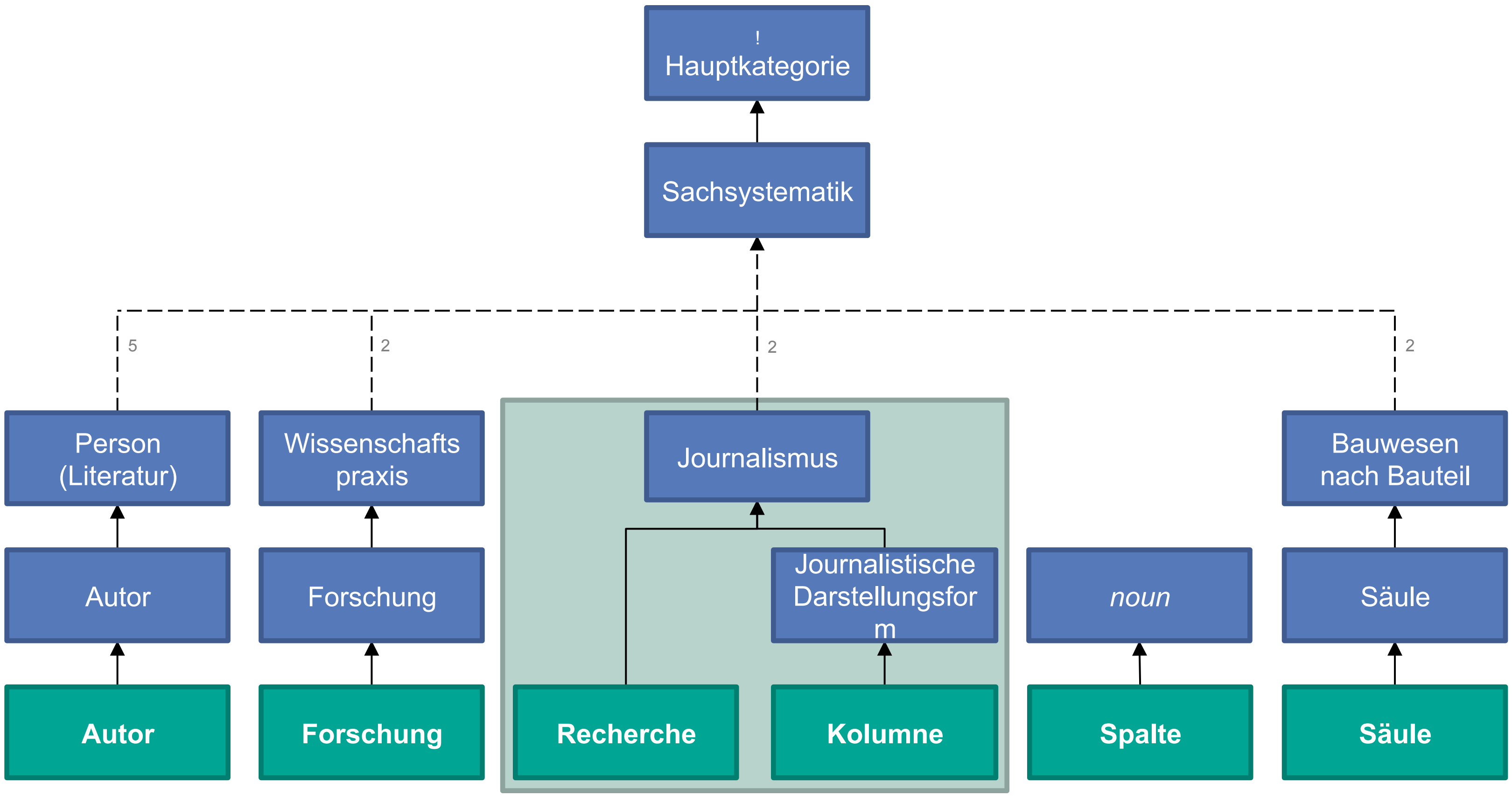


Approach 1 – Word level information

- Idea:
  - Annotate word with topic
  - Kolumne(Jornalism)
  - Säule (Building)

- Wikipedia:
  - category for each article

Approach 1 – Word level information



Motivation > Background > **Word level** > Sentence level > Evaluation > Conclusion

Approach 1 – Word level information

- Algorithm:
  - Find word in Wikipedia title
  - Determine Wikipedia category of page
  - Number of pages in category > N
    - Select category
  - Else
    - Use parent category

Approach 1 – Word level information

- Example: label nouns

	The	author	does	research	for	his	column
1.	Der	Autor	betreibt	**Forschung**	für	seine	**Säule**
	ART	Person (Literatur)	VVFIN	Wissenschafts-praxis	ADJD	PPOSAT	Bauwesen nach Bauteil
2.	Der	Autor	betreibt	**Recherche**	für	seine	**Kolumne**
	ART	Person (Literatur)	VVFIN	Journalismus	ADJD	PPOSAT	Journalismus

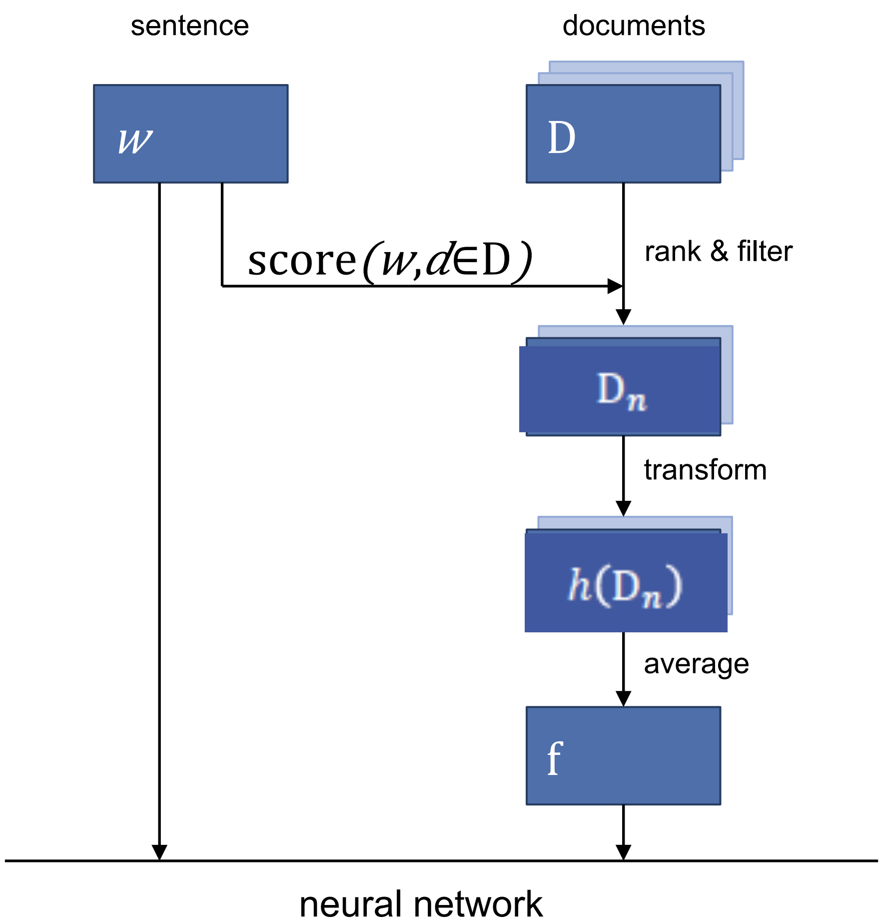
Approach 2 – Sentence level information

- Neural network can handle any input
  - Add topic information as additional input
- Model the topic of the sentence
  - Use topic model (TF-IDF/LDA/LSA/…) to represent sentence
- Use topic vector as additional input to NNLM

Approach 2 – Sentence level information

- Given translation W
- Find topic related web documents
- Represent documents in vector
  - Tf-Idf (10K dimension)
  - LSA, 300 dimension
  - LDA, 100 dimension
- Use vector in neural network

Approach 2 – sentence level information

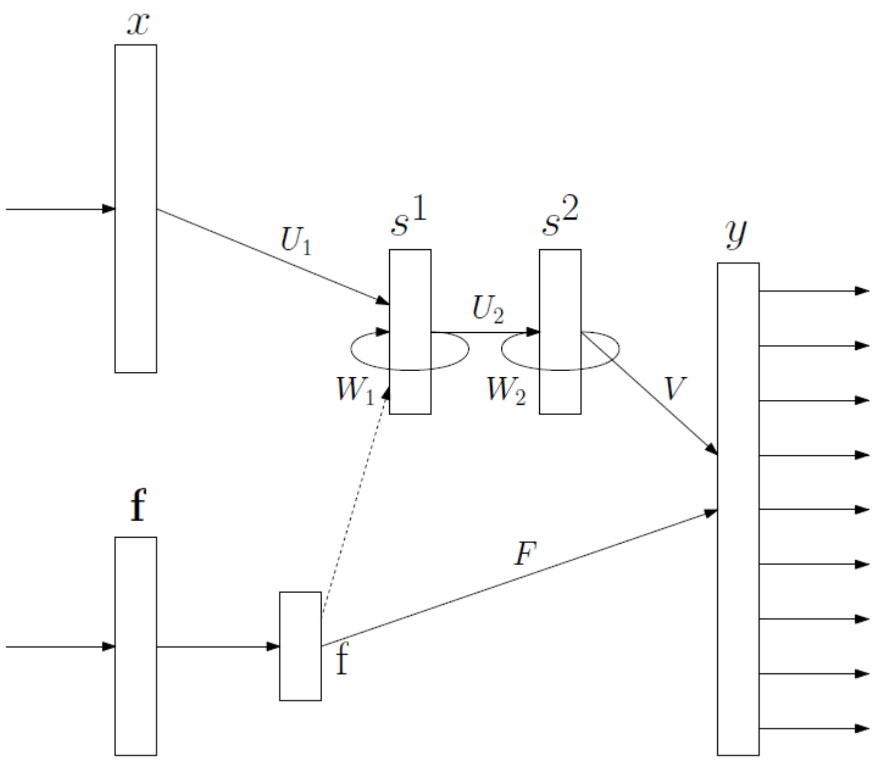


$$\mathrm{D}_n = \{\, d \in \mathrm{D} \mid \mathrm{score}(w, d) > c \,\}$$

$$|\mathrm{D}_n| = n$$

$$\mathrm{f} = \frac{\sum_{d \in \mathrm{D}_n} h(d)}{|\mathrm{D}_n|}$$

Approach 2 – sentence level information



x : word input

f : feature input

s^i: hidden layer

y : output layer

Experiment Setup

- Baseline
  - EN-ZH
    - 2010 TED, UN;
    - 3000-best list;
    - 12 features
  - EN-RO
    - WMT 2015
    - 300-best list
    - 22-23 features;

Experiment Setup

- RNN Models
  - Voc (10K (ZH) / 5K (RO)); embedding (100), LSTMs (100, 200)
  - WikiCat classes: 3000-4000
  - Number of similar documents: 10
  - SGD NLLCriterion

Evaluation – EN-ZH

Model	Devdata[BLUE]	Testdata[BLEU]
Baseline	14.70	17.02
+ Word-based WikiCat	14.89 (+0.19)	17.63 (+0.61)
+ Word-based WikiCat + POS	14.75 (+0.05)	17.81 (+0.79)
+ Sentence Wiki (Tf-idf)	14.78 (+0.08)	17.68 (+0.66)
+ Sentence Wiki 2Conn	14.74 (+0.04)	**17.81 (+0.79)**
+ Sentence ZDICT	14.91 (+0.21)	17.58 (+0.56)

Evaluation – EN-ZH

Rank	Vect	Devdata [BLEU]	Testdata [BLEU]
Baseline		14.70	17.02
TFIDF	TFIDF	14.78 (+0.08)	17.68 (+0.66)
LSA	TFIDF	14.78 (+0.08)	17.31 (+0.29)
LSA	LSA	14.83 (+0.13)	**17.80 (+0.78)**
LDA	TFIDF	14.79 (+0.09)	17.41 (+0.39)
LDA	LDA	14.79 (+0.09)	17.27 (+0.25)

Evaluation – EN-RO – Single Score

Input	Prediction	Single
Word	Word	27.88
Word, POS, 2x clusters (4F)	Word, POS, 2x clusters (4F)	28.54
+ WikiCat (N)	+ WikiCat (N)	28.71 (+0.17)
+ WikiCat (All)	+ WikiCat (All)	28.84 (+0.30)

Evaluation – EN-RO

- Word + Sentnece

Model	Conf1	Conf2	Conf3
Baseline	29.86	30.00	29.75
FRNNLM 4F	29.94	30.01	30.01
+Sentence	29.99 (+0.05)	30.19 (+0.18)	29.99 (-0.02)
+WikiCat (N)	29.90 (-0.04)	30.29 (+0.28)	30.23 (+0.22)
+WikiCat (All)	30.00 (+0.06)	30.20 (+0.19)	30.21 (+0.20)

Conclusion

- Integrate external information (from encyclopedia) into neural language models
  - use Wikipedia categories as word factors
  - Feature vector of topic-related articles for each sentence
- Improved EN-ZH by 0.79 BLEU, EN-RO by 0.2 BLEU
- Low-resource languages
- Future
  - Other web resources, topic models, network architecture
  - Use Idea on other tasks, e.g. neural machine translation

THANK YOU!