Challenges in Statistical Sign Language Translation

Christoph Schmidt, Daniel Stein, Hermann Ney
{schmidt,stein,ney}@i6.informatik.rwth-aachen.de

10.01.2011

Human Language Technology and Pattern Recognition
Lehrstuhl für Informatik 6
Computer Science Department
RWTH Aachen University, Germany

Outline

1 Corpora 3

2 Experimental Work 10

3 Challenges 17

4 Conclusions and Outlook 19

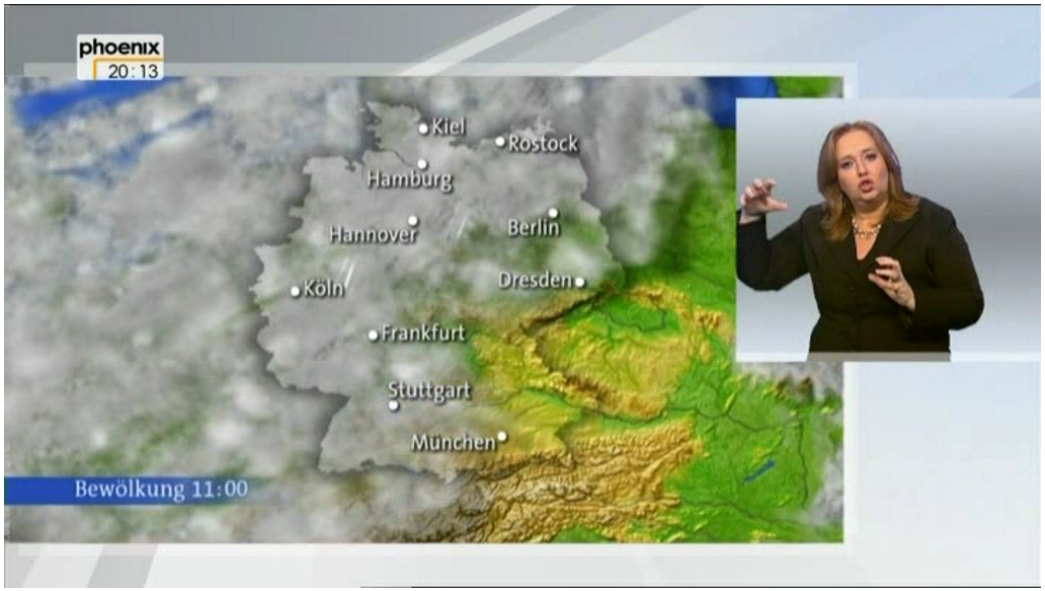
1 Corpora

Signspeak project (EU funded STREP project):

- **Better linguistic knowledge of sign languages**
- **Automatic sign language recognition**
- **Automatic sign language translation**
- **Goal: translate continuous sign language to text**

`http://www.signspeak.eu`

1.1 RWTH-Phoenix corpus



- Weather forecasts from German broadcast channel “Phoenix”
- Live interpretation into German Sign Language
- Manual annotation of glosses by deaf expert

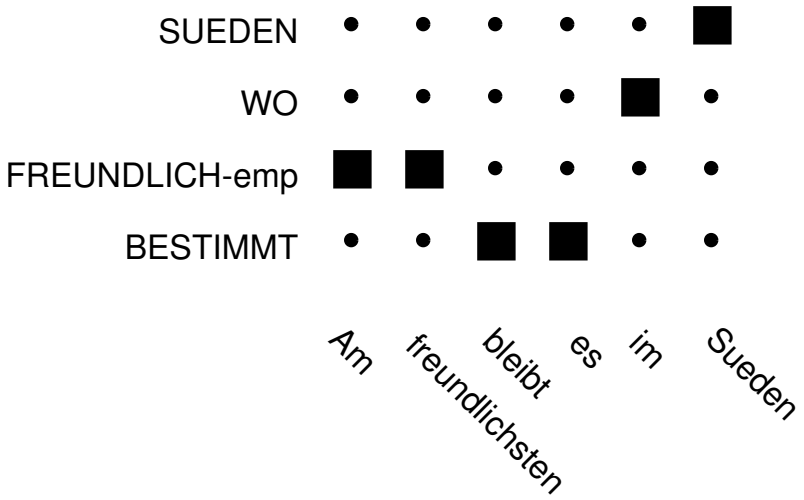
RWTH-Phoenix corpus

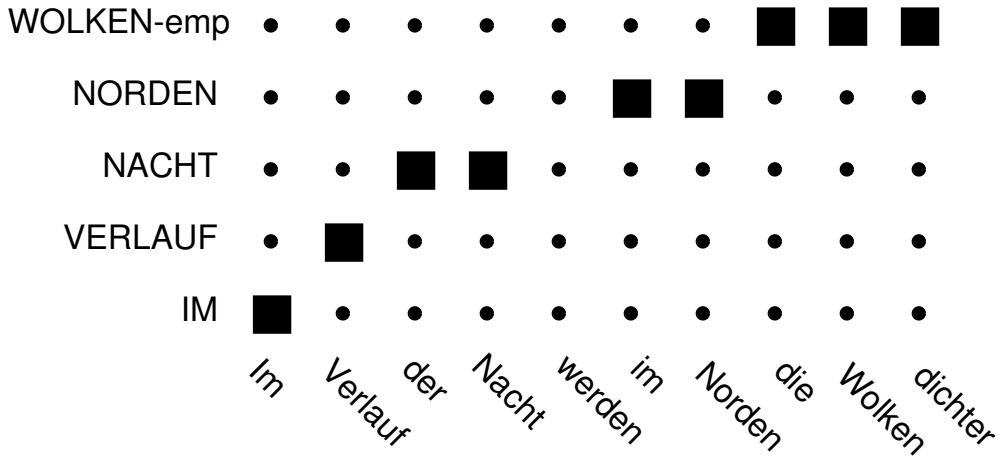
		Glosses	German
Train:	Sentences	2565	
	Running Words	31 208	41 306
	Vocabulary	1 027	1 763
	Singletons	371	641
Test:	Sentences	512	
	Running Words	6 115	8 230
	Vocabulary	570	915
	OOVs	86	133
	Trigram ppl.	51.7	22.7

Characteristics

- Narrow domain, feasible results in spite of small corpus size
- Slight bias towards German sentence structure

RWTH-Phoenix corpus





1.2 Corpus-NGT

- Data collection in Sign Language of the Netherlands
- Multiple domains: fables, cartoon paraphrases, discussions, free conversation
- Here: restriction to discussions on deafness and Deaf culture

Corpus-NGT

		Right Hand	Left Hand	Dutch
Train:	Sentences		1699	
	Running Words	8 129	4 123	15 130
	Vocabulary	1 066	773	1 695
	Singletons	481	376	840
Test:	Sentences		2.5 × 175	
	Running Words	875	496	1 815
	Vocabulary	272	181	426
	OOVs	46	39	39
	Trigram ppl.	107.0	54.6	67.5

Characteristics

- Broader domain, difficult to translate with small amount of training data
- Manual annotation of data on multiple tiers:
  left hand, right hand, non-manual signals

Corpus-NGT

right hand	MOEILIJK DOEN OVER	COMMUNICEREN PO	MET	IX HOREND	MENSEN PO
left hand	MOEILIJK DOEN	COMMUNICEREN	MET		MENSEN
Dutch	Erg veel moeite doet om te communiceren met horende mensen.				

“It is quite hard to communicate with hearing persons.”

right hand	ALS IX-1 LANG NIET			BETEKENEN EMAIL BETEKENEN	GEBAREN IX-1 PO	
left hand		NIET HEEN CLUBHUIS TOE BETEKENEN				GEBAREN IX-1 PO
Dutch	als je lang niet naar het clubhuis gaat , weet je het gebaar voor het woord e-mail bijvoorbeeld niet .					

**“If you haven’t been to the club house for some time,
you won’t know the sign for the word ‘email’ ”**

2 Experimental Work

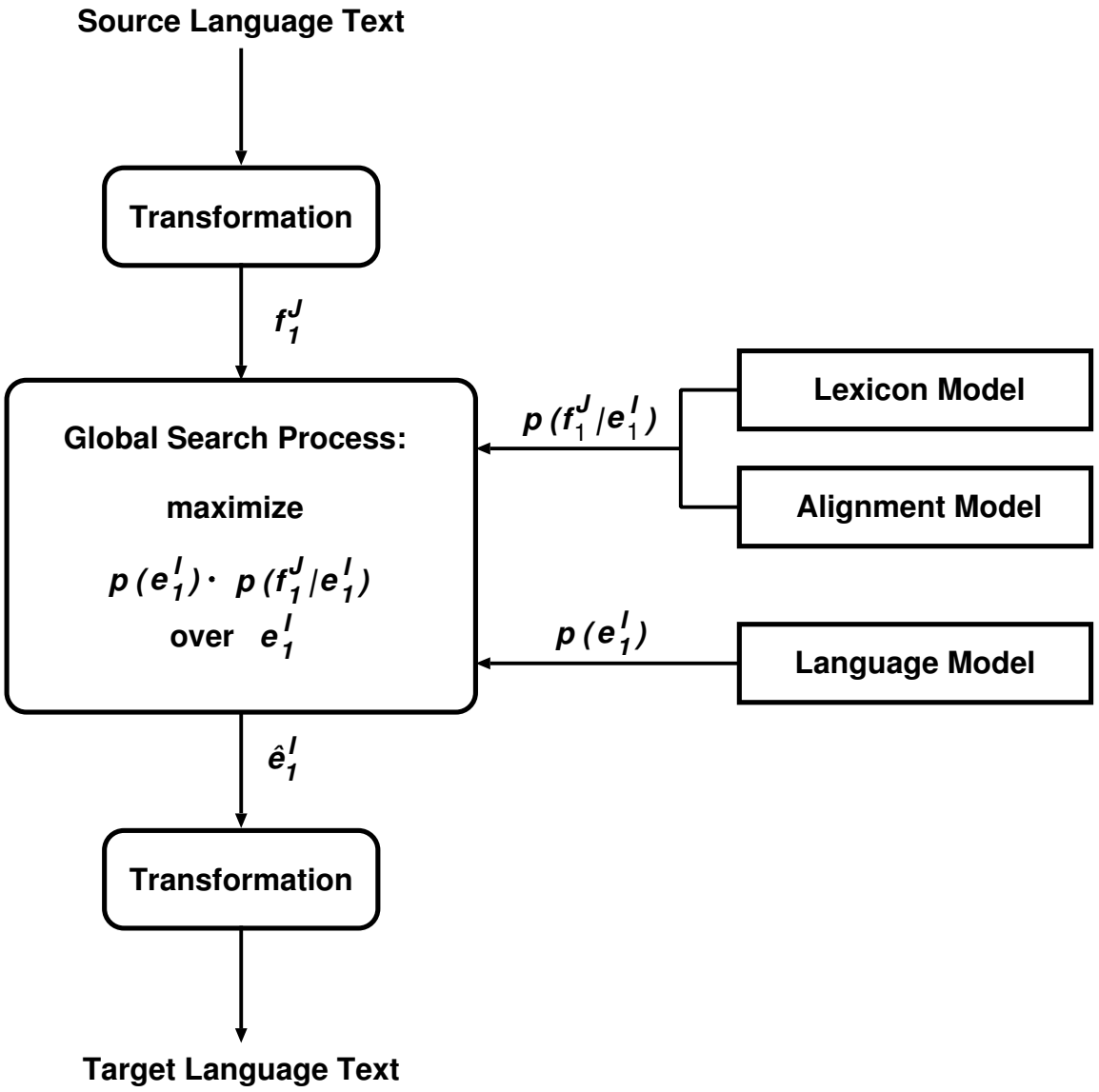
Statistical machine translation (SMT):

- **Using statistical models to translate text (“most probable translation”)**
- **Models are trained on bilingual corpora**
- **State of the art in spoken language translation**

Advantage of statistical MT:

- **Language independence**
- **Setting up a translation system for a new language pair by providing a bilingual corpus**

Statistical MT



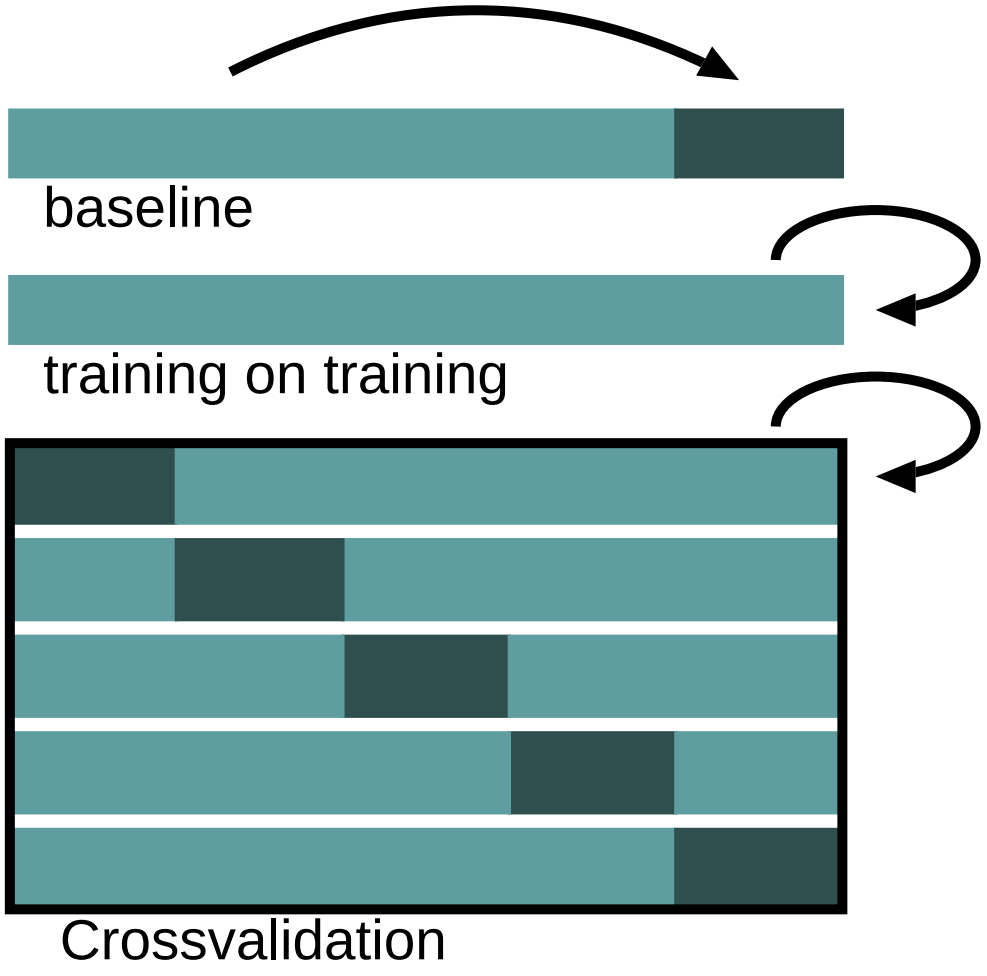
Methods

- Different alignment merging strategies
- Different paradigms: phrase-based vs. hierarchical translation
- Adaptation of standard methods to small corpora: cross-validation
- Advanced lexicon models (DWL, triplet model)

Pre/Postprocessing

- Use world knowledge: categories ($weekday, $month, $number, $ordinal)
- Sentence end markers on gloss side
- Use linguistic knowledge of spoken language: compound splitting (Morphisto)

Methods: “Crossvalidation”



RWTH-Phoenix Corpus

Phrase-based system

	BLEU	TER
baseline	14.2	79.3
+ categories	15.4	77.0
+ compound splitting	15.9	75.8
+ advanced lexicon models	16.3	74.3

Hierarchical system

	BLEU	TER
baseline (incl. categories)	15.5	84.5
+ sentence end markers	16.3	76.1
+ compound splitting	16.1	74.1

Corpus-NGT

How to evaluate multiple streams?

right hand	ALS IX-1 LANG NIET	BETEKENEN EMAIL BETEKENEN GEBAREN IX-1 PO
left hand	NIET HEEN CLUBHUIS TOE BETEKENEN	GEBAREN IX-1 PO
Dutch	als je lang niet naar het clubhuis gaat , weet je het gebaar voor het woord e-mail bijvoorbeeld niet .	

Phrase-based system

	BLEU	TER
Right hand	**4.4**	**104.8**
Active hand	**5.1**	**90.4**
Merge hands	**2.7**	**82.3**

⇒ **Results still unstable**

Translation directions

RWTH-Phoenix	Phrase-Based		Hierarchical	
	BLEU	TER	BLEU	TER
DE⇒GL	16.3	74.3	16.1	74.1
GL⇒DE	25.4	62.9	25.0	66.5

NGT-Corpus	BLEU	TER
NL⇒GL	5.1	90.4
GL⇒NL	10.7	78.3

- Translation direction to glosses seems to be more difficult

3 Challenges

General challenges:

- **Data size of available sign language corpora (thousands of sentences vs. millions of sentences in spoken languages)**
- **Costly annotation of sign language corpora**
- **Different means of communication: hand motion, body posture, facial expression, etc.**
- **Paraphrases vs. literal translation**

Challenges

Challenges of RWTH-Phoenix:

- **Omissions in the interpretation (generalization)**
- **Pointing to the map**

$\Rightarrow$ **Information mismatch**

Challenges of Corpus-NGT

- **Corpus too small for domain**
- **Casual conversation style: hesitations, partial sentences**
- **Translations make use of facial expressions (“I totally agree”)**
- **Handling of multiple tiers (hands, head-shakes)**

But: more natural than Phoenix corpus

4 Conclusions and Outlook

Conclusions

- Application of SMT to sign languages is possible
- Main obstacle: small corpus sizes
- Adaptation of methods to small corpora
- Open question: multiple communication channels

Outlook

- Head-shakes as negation markers
- Alleviate information mismatch by improving corpora
- Process multiple communication channels in parallel

Thank you for your attention

Christoph Schmidt,Daniel Stein,Hermann Ney

schmidt@i6.informatik.rwth-aachen.de

http://www-i6.informatik.rwth-aachen.de/