Centre for
Data Analytics

Insight

IRIS - English-Irish Translation System

Mihael Arcan, Unit for Natural Language Processing of the Insight Centre for Data Analytics at the National University of Ireland, Galway

DCU NUI Galway Oé Gaillimh UCC UCD DUBLIN sfi

Introduction

- about me, Natural Language Processing Unit, my study and motivation, ...

- Statistical Machine Translation (SMT)
  - SMT training, word/phrase alignments, ambiguity, examples, evaluation

- IRIS - English-Irish Translation System
  - used resources, evaluation, future work

about me*

- PhD Student at Insight Centre for Data Analytics @ NUI Galway
  - supervised by Dr. Paul Buitelaar
- studied German Language (Diploma study) at the University of Ljubljana, Slovenia
- Masters degree in Computational Linguistics at the Ruhr University in Bochum, Germany

http://nlp.insight-centre.org/people/members/mihael-arcan/

Unit for Natural Language Processing

http://nlp.insight-centre.org/

Reseach Topics in UNLP

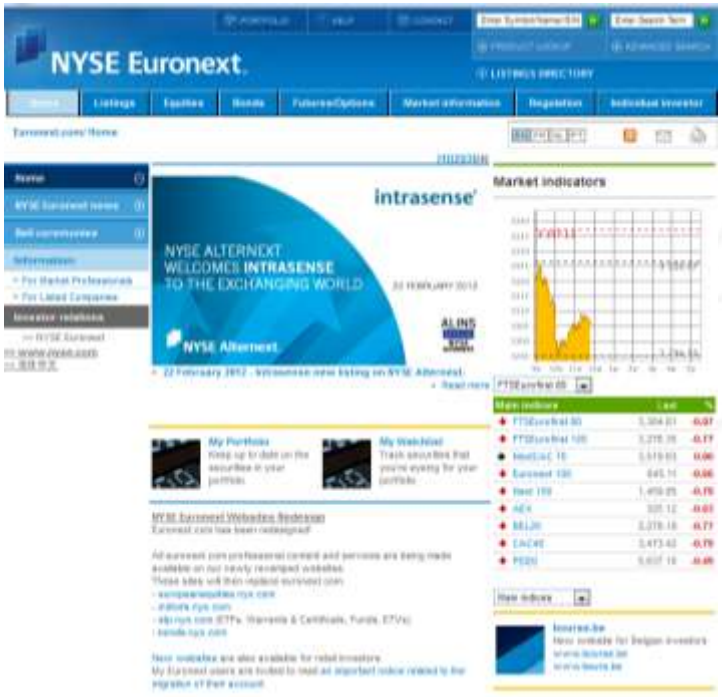
- Entity Linking
- Expertise Mining
- Linguistic Linked Data
- Content-based Linked Data Summarisation
- Semantic Similarity and Relatedness
- Suggestion Extraction
- Taxonomy Construction
- Term Translation

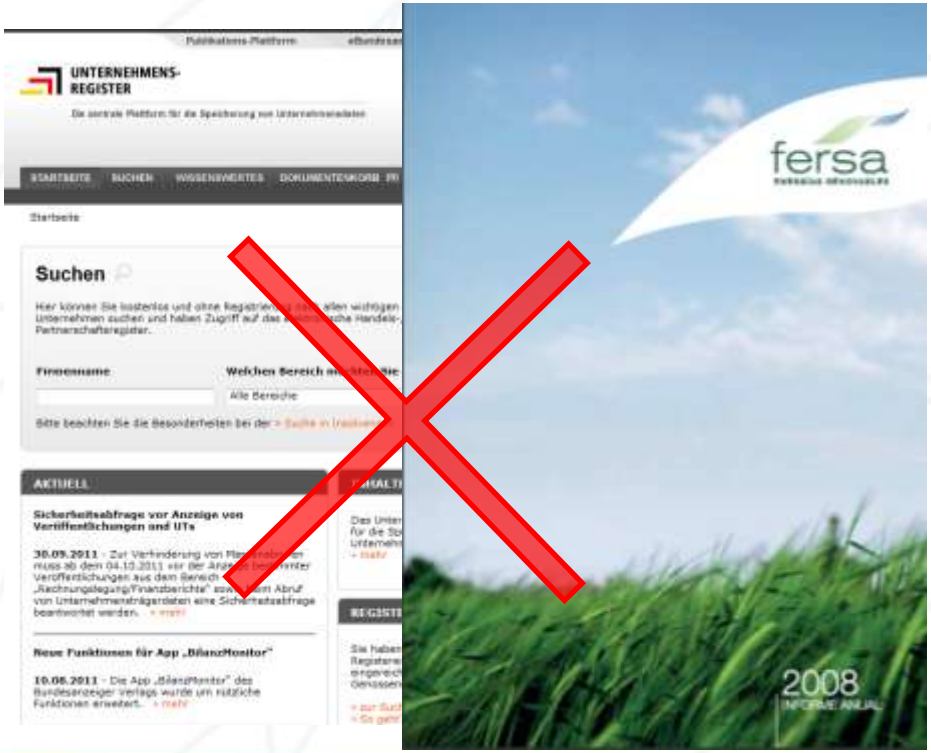
http://nlp.insight-centre.org/

Motivation of my Study

Business Information query in English

equity@en

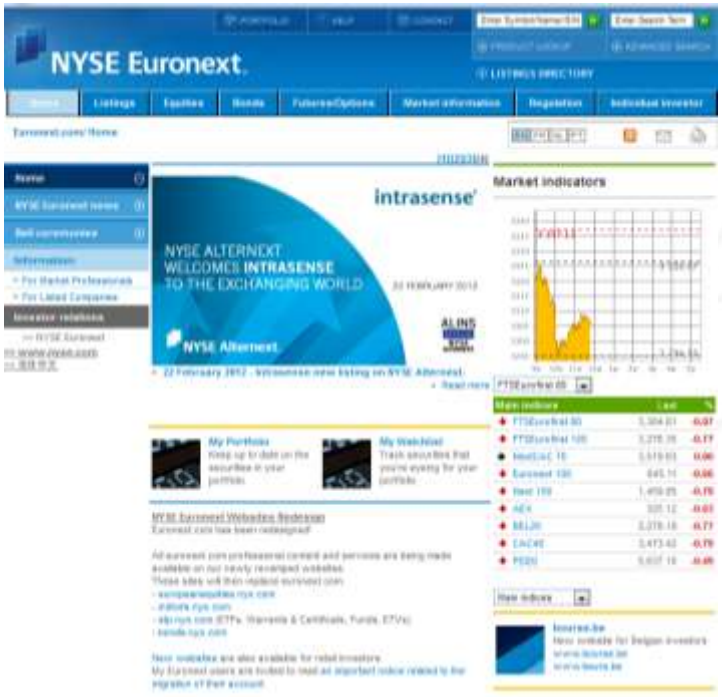


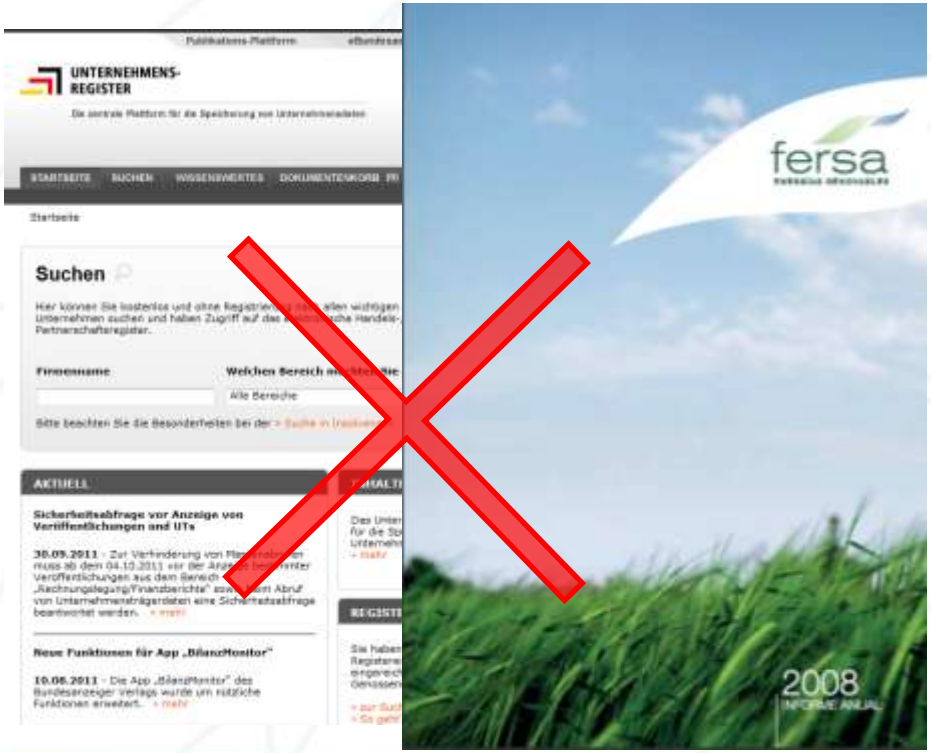


Motivation of my Study

Business Information query in English

equity@en -> Google Translate -> Gerechtigkeit@de

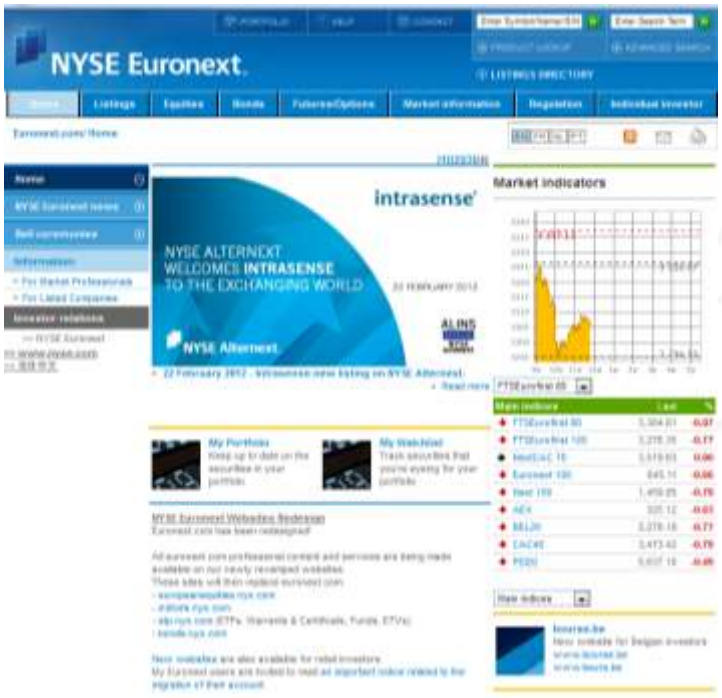




Motivation of my Study

Business Information query in English

equity@en -> domain aware SMT-> Eigenkapital@de



Issue 1 with SMT (in Term Translation)

Source text:

bartonellosis(1)

Reference text:

bartonellose

Target text (generic translation model)

bartonellosis(2)

bartonellosis(3)

Out-of-Vocabulary (OOV) Problem

Target text (domain-specific model)

bartonellosis

(1) ICD (International Classification of Diseases) ontology
(2) .../tetra_old/, general
(3) Google Translate, 2.6.'15

Issue 2 with SMT (in Term Translation)

Source text:

cash flow hedges

Reference text:

Absicherungen von Zahlungsströmen

Target text (generic model)

Cashflow-Hedges (€)(1)

Cashflow Hecken(2)

Out-of-Domain Translation

Target text (domain-specific model3)

Absicherungen von Zahlungsströmen (€)

1) Google Translate, 3.5.'15
(2) .../tetra_old/, general
((3) http://server1.nlp.insight-centre.org/tetra_old/, financial

Introduction

- Natural Language Processing Unit, about me, my study, motivation, ...
- Statistical Machine Translation (SMT)
  - SMT training, word/phrase alignments, ambiguity, examples, evaluation
- IRIS - English-Irish Translation System
  - used resources, evaluation, future work

Models in Statistical Machine Translation

- Translation Model
  - lexical correspondence between languages

fixed asset | anlagevermögen | 0.003 0.003 0.029 0.102

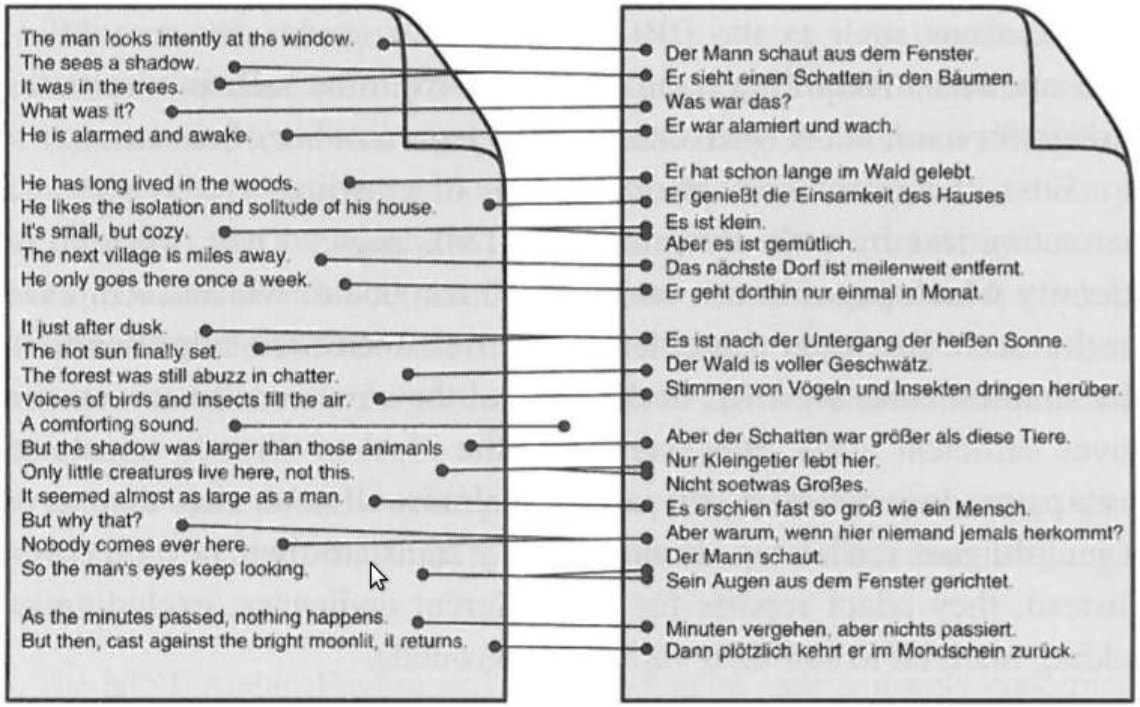
- Language Model
  - takes care of fluency (and lexical choice) in the target language

-4.868038	anlagevermögen	-0.1317768

Training Data (Parallel Corpora) for SMT

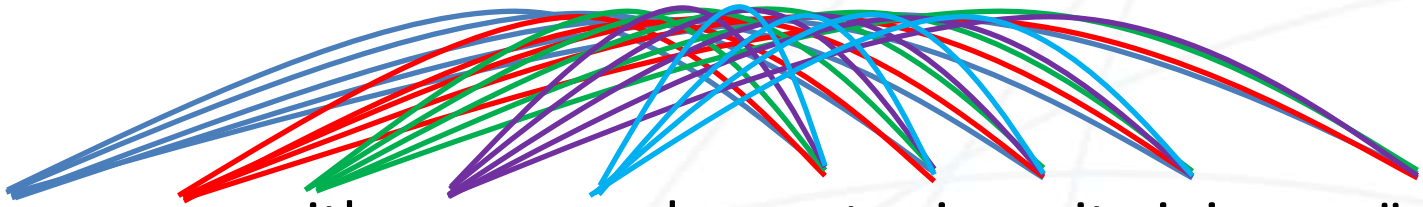
Source document

Target document



Training Data for SMT

- Sentence aligned parallel data

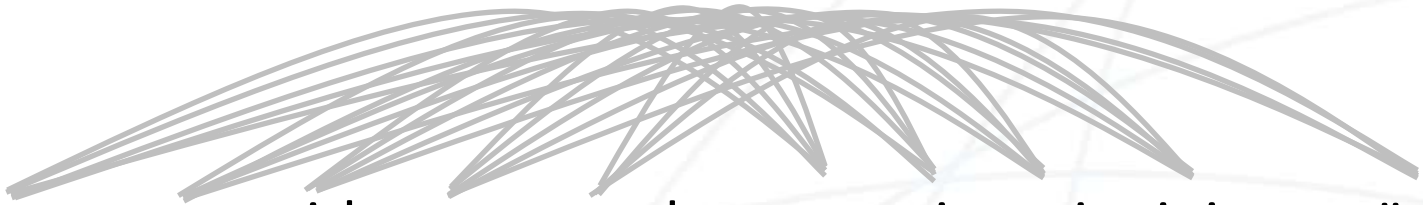


sentence one with some words
sentence two with more words
…

satz eins mit einigen wörtern
satz zwei mit weiteren wörtern
…

Training Data for SMT

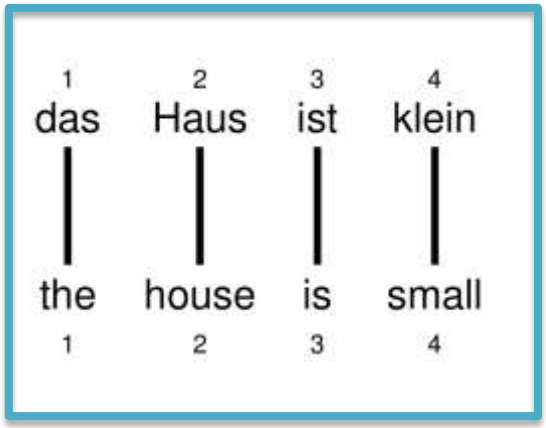
- Sentence aligned parallel data → Word Alignment

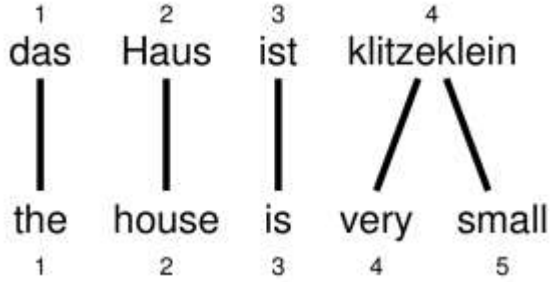


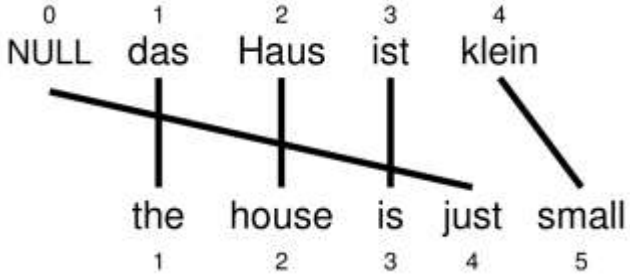
sentence one with some words
sentence two with more words
…

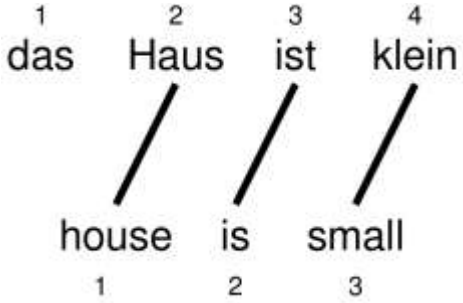
satz eins mit einigen wörtern
satz zwei mit weiteren wörtern
…

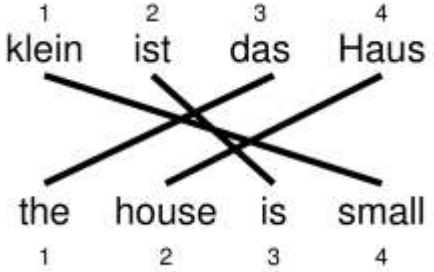
Word Alignment Scenarios



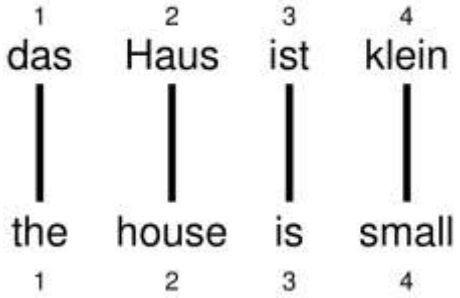


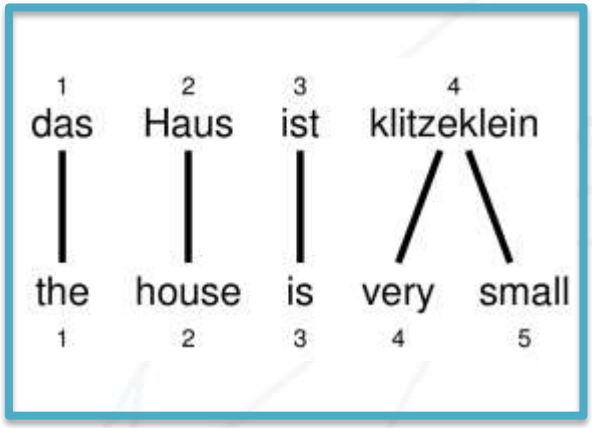


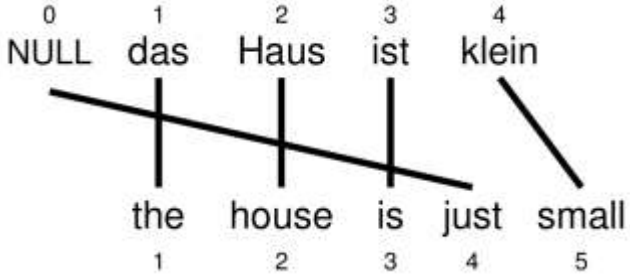


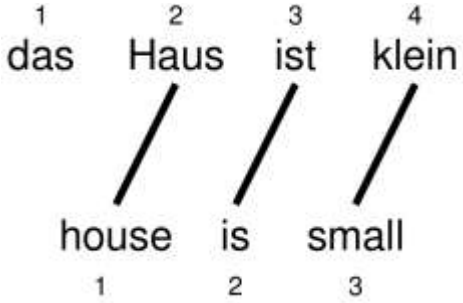


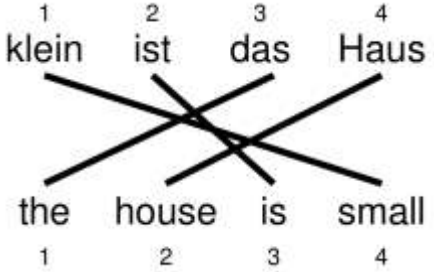
Word Alignment Scenarios



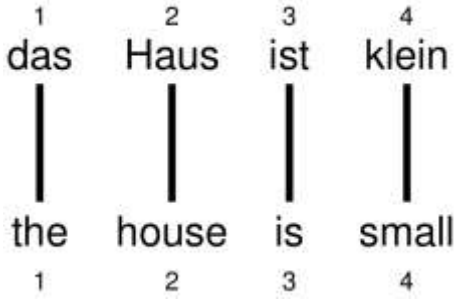


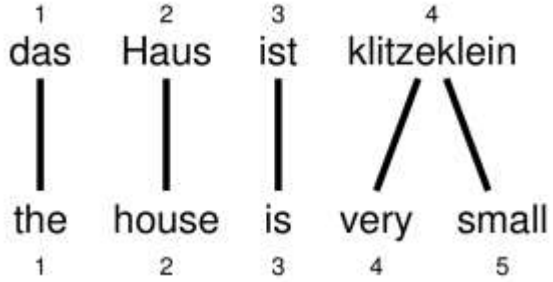


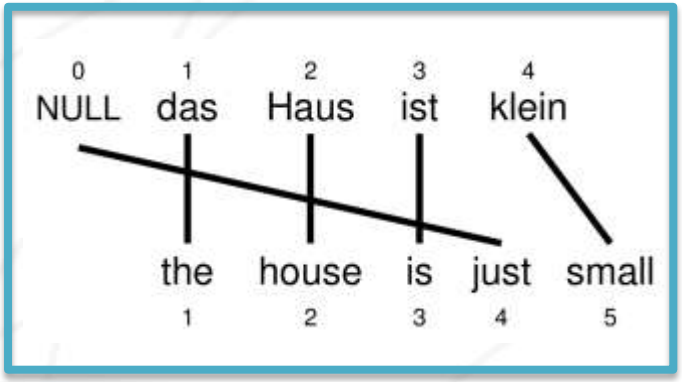


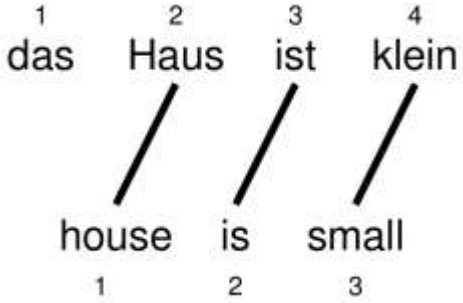


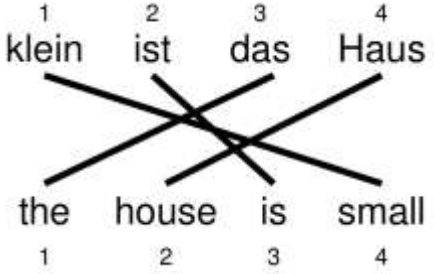
Word Alignment Scenarios



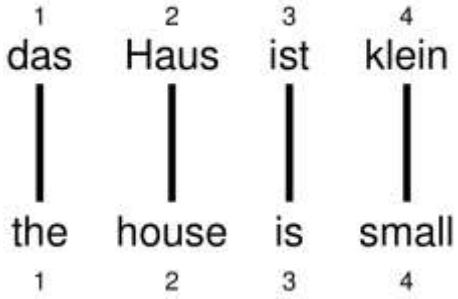


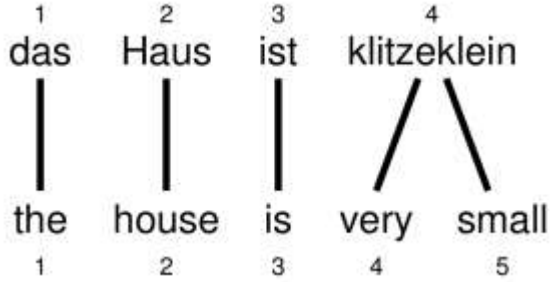


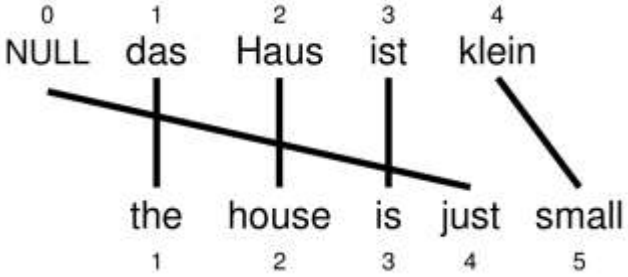


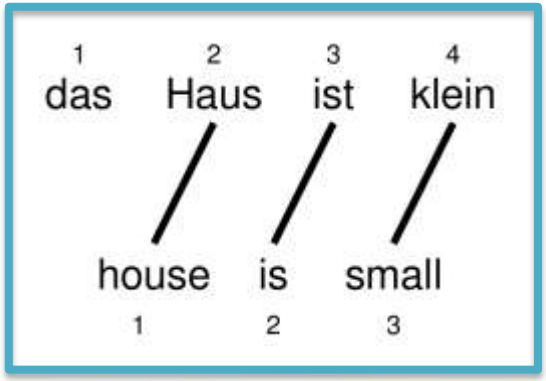


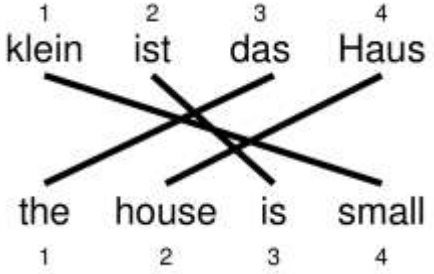
Word Alignment Scenarios



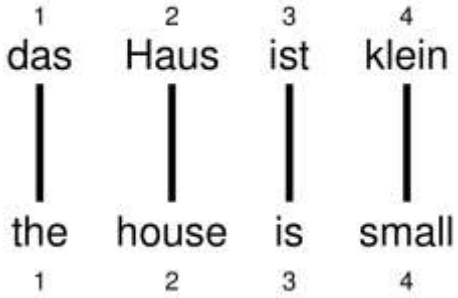


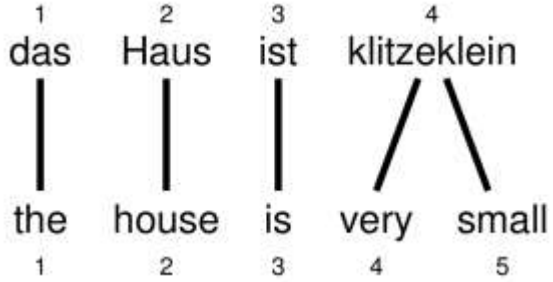


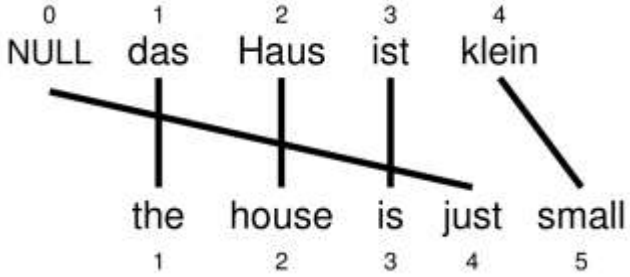


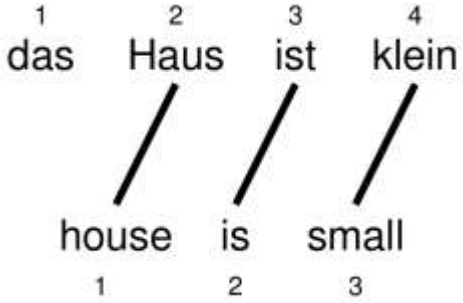


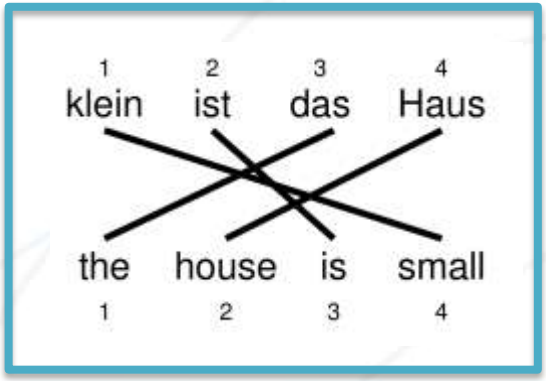
Word Alignment Scenarios











Word Alignment with IBM Models

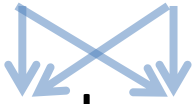
Models	Function of the model
IBM Model 1	lexical translation
IBM Model 2	adds absolute reordering model
IBM Model 3	adds fertility model
IBM Model 4	adds relative alignment model
IBM Model 5	fixes deficiency

IBM Model 1 – Practical Session

Training data:

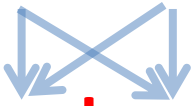
Source Document	Target Document
das Haus	the house
das Buch	the book
ein Buch	a book

das haus



the house

das **buch**

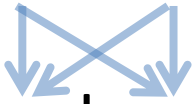


the book

ein **buch**

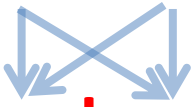
a book

das haus



the house

das **buch**



the book

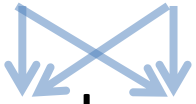
ein **buch**

a book

e	f	Initial	1st iter.	2nd iter.	3rd iter.	...	Final
the	das	0.25					1
book	das	0.25					0
house	das	0.25					0
the	buch	0.25					0
book	buch	0.25					1
a	buch	0.25					0
book	ein	0.25					0
a	ein	0.25					1
the	haus	0.25					0
house	haus	0.25					1

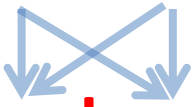
Learning Phase

das haus



the house

das **buch**



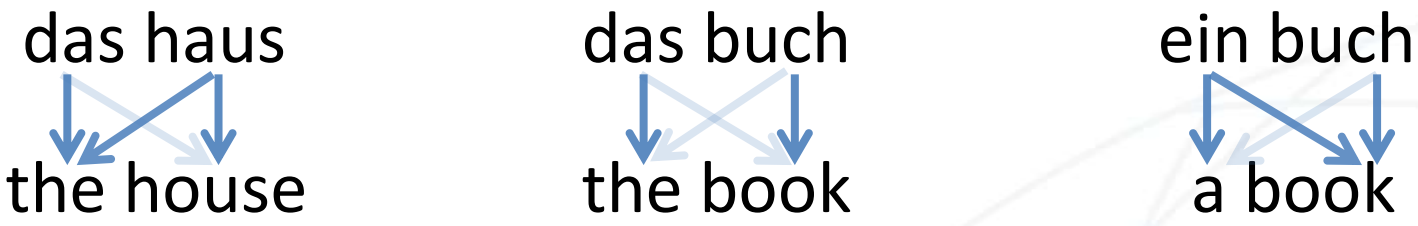
the book

ein **buch**

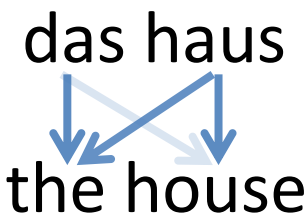
a book

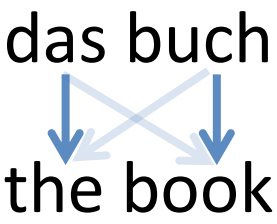
e	f	Initial	1st iter.	2nd iter.	3rd iter.	...	Final
the	das	0.25					1
book	das	0.25					0
house	das	0.25					0
the	buch	0.25					0
book	buch	0.25					1
a	buch	0.25					0
book	ein	0.25					0
a	ein	0.25					1
the	haus	0.25					0
house	haus	0.25					1

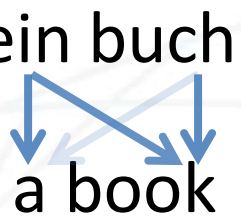
Learning Phase



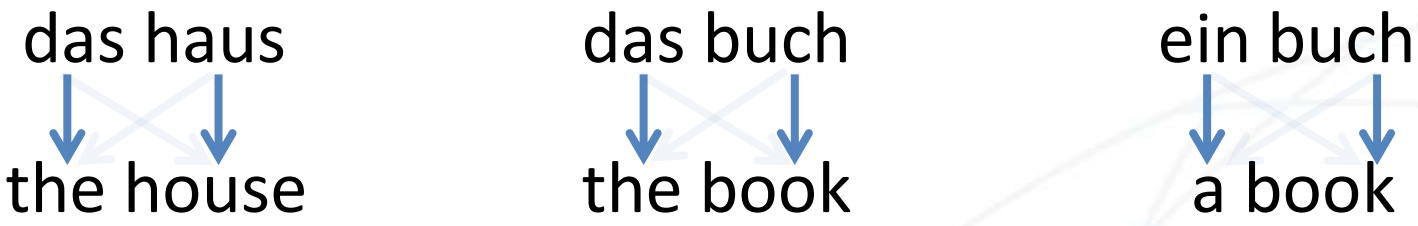
e	f	Initial	1st iter.	2nd iter.	3rd iter.	...	Final
the	das	0.25	**0.50**				1
book	das	0.25	0.25				0
house	das	0.25	0.25				0
the	buch	0.25	0.25				0
book	buch	0.25	**0.50**				1
a	buch	0.25	0.25				0
book	ein	0.25	**0.50**				0
a	ein	0.25	**0.50**				1
the	haus	0.25	**0.50**				0
house	haus	0.25	**0.50**				1







e	f	Initial	1st iter.	2nd iter.	3rd iter.	...	Final
the	das	0.25	0.50	**0.6364**			1
book	das	0.25	0.25	0.1818			0
house	das	0.25	0.25	0.1818			0
the	buch	0.25	0.25	0.1818			0
book	buch	0.25	0.50	**0.6364**			1
a	buch	0.25	0.25	0.1818			0
book	ein	0.25	0.50	0.4286			0
a	ein	0.25	0.50	**0.5714**			1
the	haus	0.25	0.50	0.4286			0
house	haus	0.25	0.50	**0.5714**			1



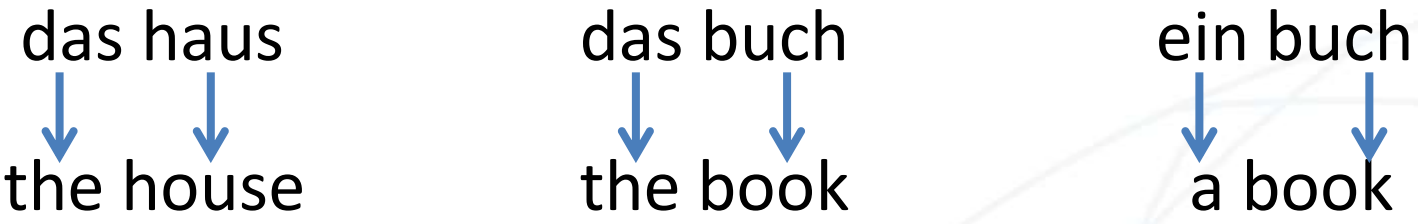
e	f	Initial	1st iter.	2nd iter.	3rd iter.	...	Final
the	das	0.25	0.50	0.6364	**0.7479**		1
book	das	0.25	0.25	0.1818	0.1208		0
house	das	0.25	0.25	0.1818	0.1313		0
the	buch	0.25	0.25	0.1818	0.1208		0
book	buch	0.25	0.50	0.6364	**0.7479**		1
a	buch	0.25	0.25	0.1818	0.1313		0
book	ein	0.25	0.50	0.4286	0.3466		0
a	ein	0.25	0.50	0.5714	**0.6534**		1
the	haus	0.25	0.50	0.4286	0.3466		0
house	haus	0.25	0.50	0.5714	**0.6534**		1

das haus → the house

das buch → the book

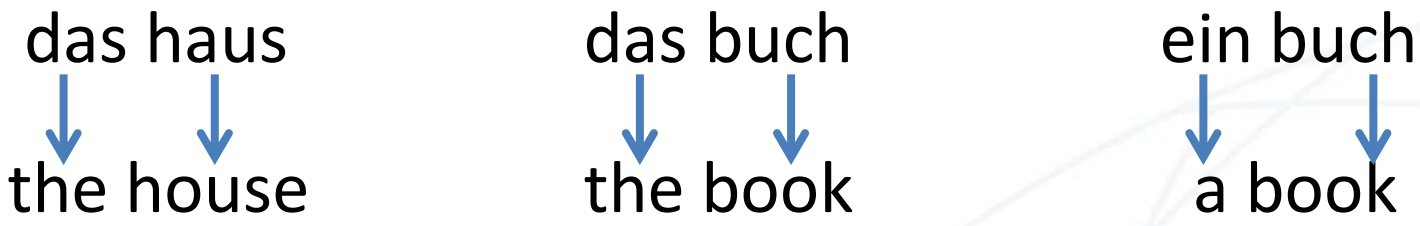
ein buch → a book

e	f	Initial	1st iter.	2nd iter.	3rd iter.	...	Final
the	das	0.25	0.50	0.6364	0.7479	...	**1**
book	das	0.25	0.25	0.1818	0.1208	...	0
house	das	0.25	0.25	0.1818	0.1313	...	0
the	buch	0.25	0.25	0.1818	0.1208	...	0
book	buch	0.25	0.50	0.6364	0.7479	...	**1**
a	buch	0.25	0.25	0.1818	0.1313	...	0
book	ein	0.25	0.50	0.4286	0.3466	...	0
a	ein	0.25	0.50	0.5714	0.6534	...	**1**
the	haus	0.25	0.50	0.4286	0.3466	...	0
house	haus	0.25	0.50	0.5714	0.6534	...	**1**



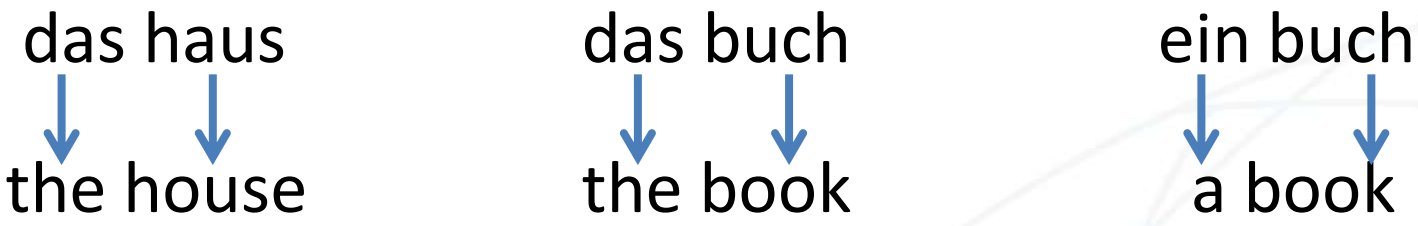
Lexical (word) probabilities (after 10 iterations):

- buch
  - book 0.9933
  - a 0.0046
  - the 0.0020
- haus
  - house 0.9172
  - the 0.0827
- das
  - the 0.9933
  - house 0.0046
  - book 0.0020
- ein
  - a 0.9172
  - book 0.0827



Decoding (translating) using the lexical probabilities:

- ein buch
  - a book 0.25
  - book book 0.01
- das buch
  - the book 0.25
- das haus
  - the house 0.25
  - the the 0.01



Decoding (translating) using the lexical probabilities:

- ein buch
  - a book 0.25
  - book book 0.01
- das buch
  - the book 0.25
- das haus
  - the house 0.25
  - the the 0.01
- ein haus
  - a house 0.25
  - book house 0.01

Language Ambiguity in SMT

Source language	Target language
freundliche[1] bank[2]	friendly[1] bank[2]
gemütliche[3] bank[4]	cosy[3] bench[4]
freundliche[1]	friendly[1]
gemütliche[5]	cosy[5]
schlechte[6]	bad[6]

schlechte[6] bank[2]

bad[6] bench[4] 0.1239

bad[6] bank[2] 0.1239

Generic Models in SMT

Source language	Target language
freundliche[1] bank[2]	friendly[1] bank[2]
gemütliche[3] bank[4]	cosy[3] bench[4]
freundliche[1]	friendly[1]
gemütliche[5]	cosy[5]
schlechte[6]	bad[6]
grüne[7] bank[4]	green[7] bench[4]

schlechte[6] bank[2]

bad[6] bench[4] 0.1918 ✗

bad[6] bank[2] 0.0581

Domain aware Models in SMT

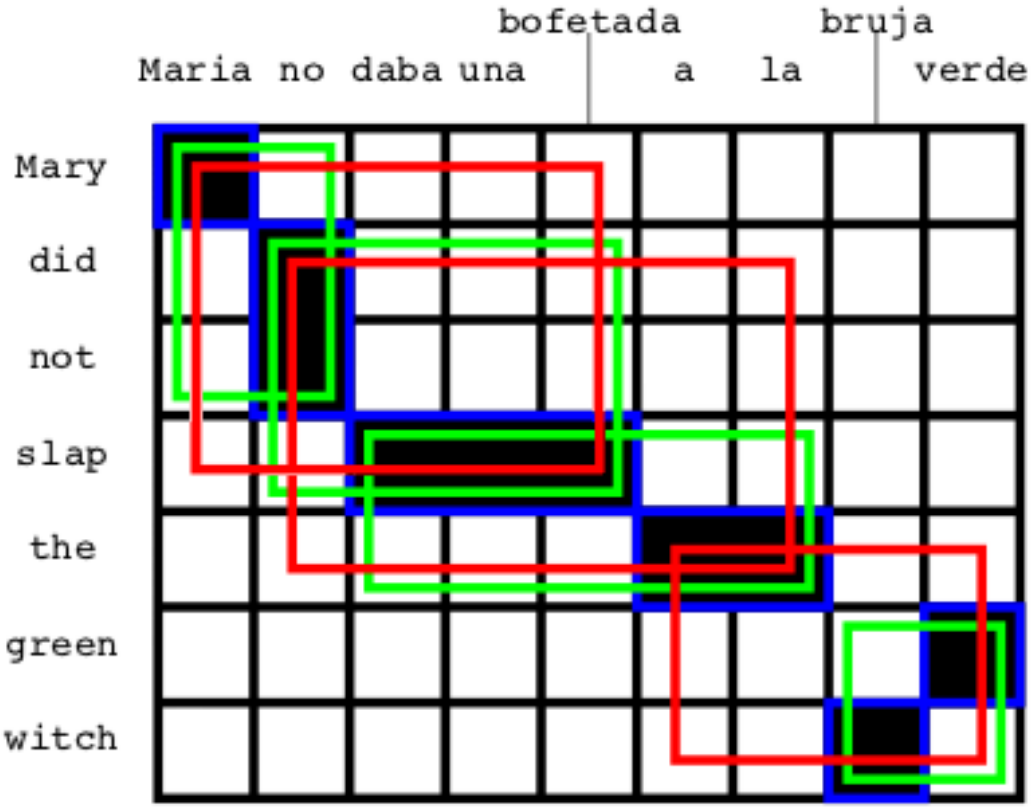
Source language	Target language
freundliche[1] bank[2]	friendly[1] bank[2]
gemütliche[3] bank[4]	cosy[3] bench[4]
freundliche[1]	friendly[1]
gemütliche[5]	cosy[5]
schlechte[6]	bad[6]
multinationale[7] bank[2]	multinational[7] bank[2]

schlechte[6] bank[2]

bad[6] bank[2] 0.1918

bad[6] bench[4] 0.0581

From Word to Phrase Based SMT



- Maria, Mary
- no, did not
- daba una bofetada, slap
- a la, the
- bruja, witch
- verde, green

- Maria no, Mary did not
- no daba una bofetada, did not slap
- daba una bofetada a la, slap the
- bruja verde, green witch

- Maria no daba una bofetada, Mary did not slap
- no daba una bofetada a la, did not slap the
- a la bruja verde, the green witch

Generic Models in SMT

Source language	Target language
freundliche[1] bank[2]	friendly[1] bank[2]
gemütliche[3] bank[4]	cosy[3] bench[4]
freundliche[1]	friendly[1]
gemütliche[5]	cosy[5]
schlechte[6]	bad[6]
grüne[7] bank[4]	green[7] bench[4]

freundliche[1] bank[2]

friendly[1] bench[4] 0.1576 ✗

friendly[1] bank[2] 0.0581

Generic Models in SMT

Source language	Target language
freundliche_bank[1]	friendly_bank[1]
gemütliche_bank[1]	cosy_bench[2]
freundliche[3]	friendly[3]
gemütliche[4]	cosy[4]
schlechte[5]	bad[5]
grüne_bank[6]	green_bench[6]

freundliche_bank[1]

friendly_bank[1] 1.0

Why are phrases better?

Prime Minister Ayrault said: "It's incredible that an **allied country** like the United States at this point goes as far as spying on private communications that have no strategic justification, no justification **on the basis of national** defence.“

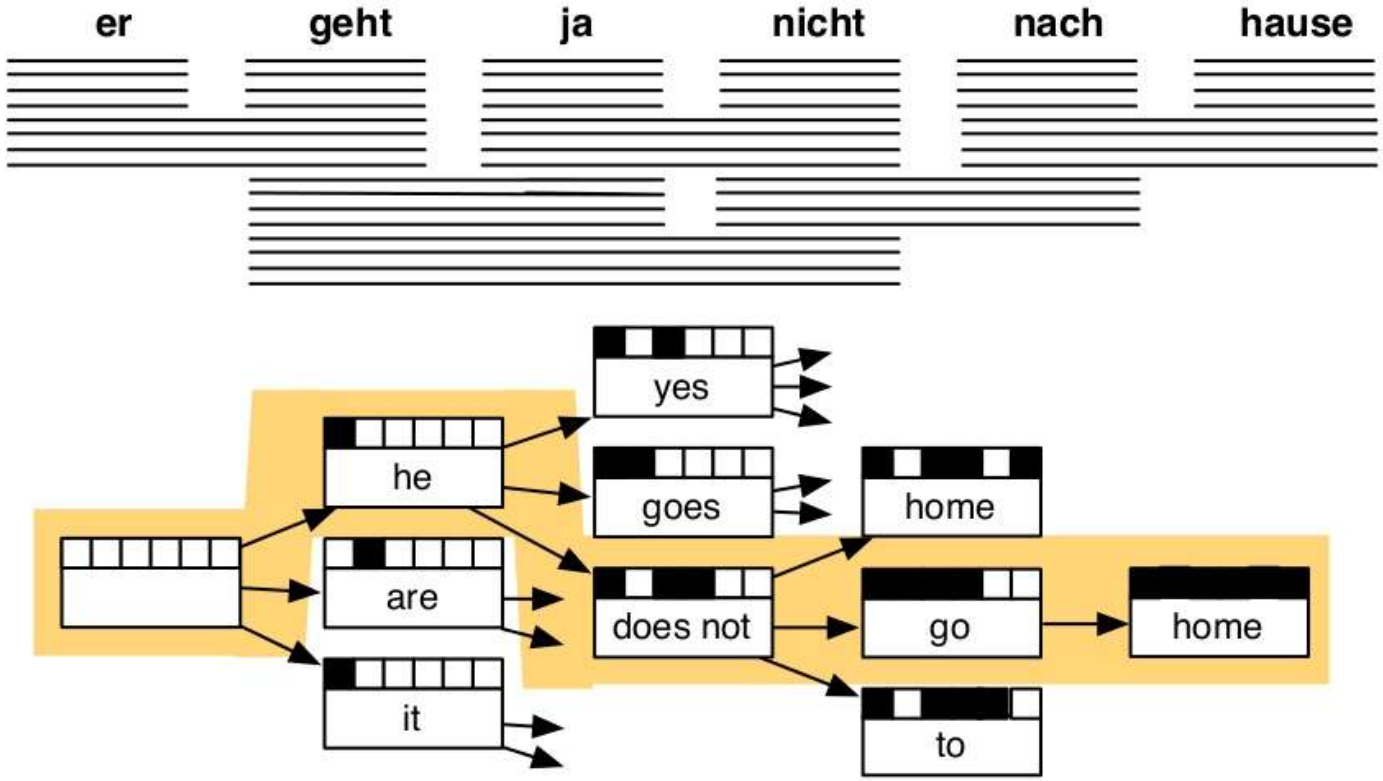
Premierminister | ayrault | sagte: |"es | ist unglaublich |, dass eine | **verbündete | Land** wie die | Vereinigten Staaten | an diesem Punkt | geht so | weit wie | Spionage | auf private | Mitteilungen | , dass | keine strategische |Gründe | , keine | Begründung | **auf der Grundlage der nationalen** | Verteidigung. |"

Decoding

Lexical probabilities:

- buch
  - book 0.9933
  - a 0.0046
  - the 0.0020
- haus
  - house 0.9172
  - the 0.0827
- das
  - the 0.9933
  - house 0.0046
  - book 0.0020
- ein
  - a 0.9172
  - book 0.0827

Decoding - Finding the best path

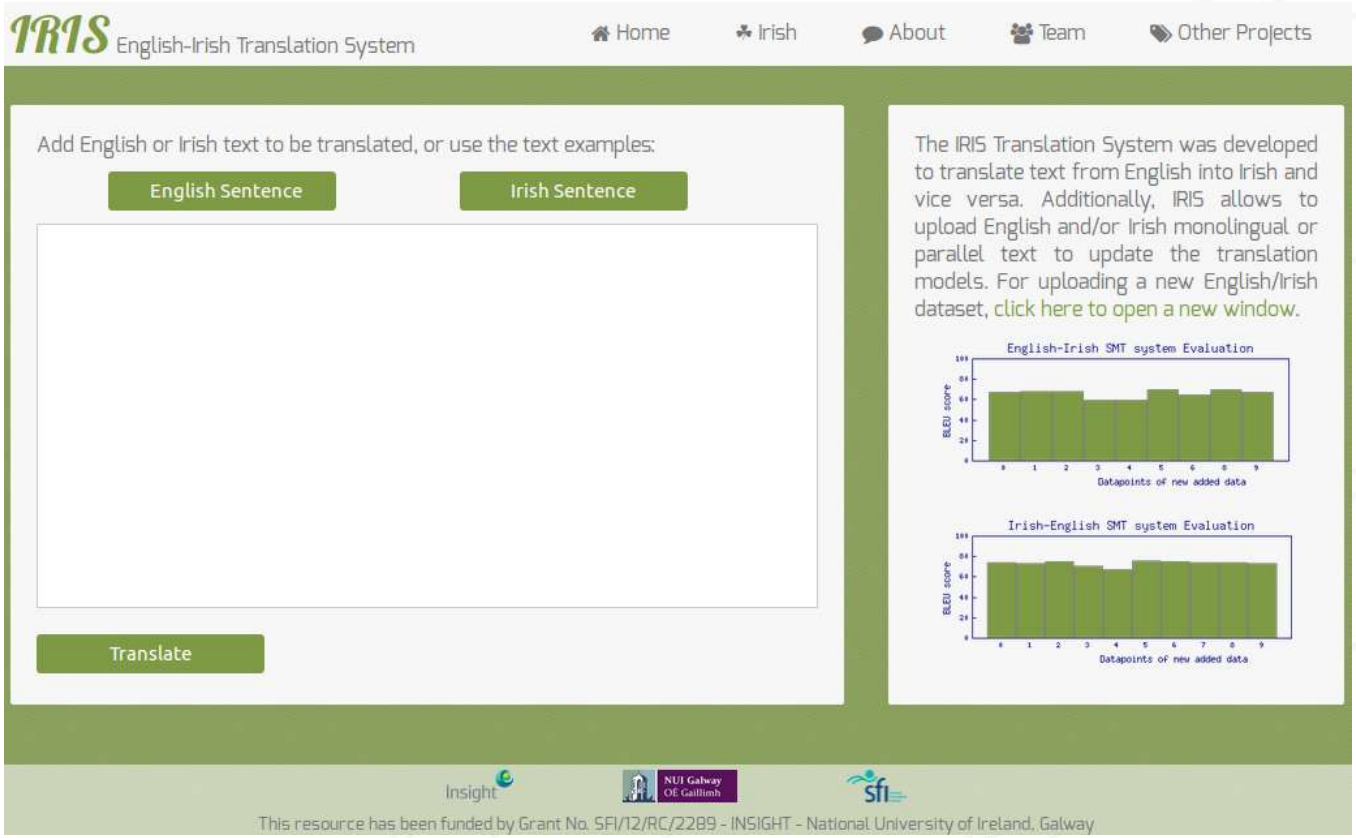


Best translation = probability of Translation Model * Language Model

Introduction

- Natural Language Processing Unit, about me, my study, motivation, ...
- Statistical Machine Translation (SMT)
  - SMT training, word/phrase alignments, ambiguity, examples, evaluation
- **IRIS - English-Irish Translation System**
  - **used resources, evaluation, future work**

IRIS English-Irish Translation System*



http://server1.nlp.insight-centre.org/iris/

IRIS English-Irish Translation System*

- it uses Moses Translation Toolkit
- translates from English into Irish and Irish into English (with optional translations)
- Irish interface
- build on publically available data
- used in UNLP/Insight projects (Kennys Bookshop, DBpedia as Gaeilge, OTTO)
- alows users to add new data to the system

http://server1.nlp.insight-centre.org/iris/

Resources for IRIS

Resource	# of lines	# English words	# Irish words	BLEU	
DGT.en-ga	36,275	864,373	950,500	27.64	36.70
EUbookshop.en-ga	121,042	2,606,607	2,704,091	49.33	56.50
EUconst.en-ga	6,267	125,553	126,355	49.92	57.62
focal_en_ga...	213,683	414,730	440,228	48.57	59.14
GNOME.en-ga	75,051	288,916	297,882	53.23	60.80
irish-legislation...	132,314	2,691,928	2,792,595	52.72	60.30
KDE4.en-ga	110,138	439,273	523,614	53.83	59.72
news.2007.shuffled*	3,782,548	/	90,490,396		60.14
Ubuntu.en-ga	191	1,038	1,103	53.61	61.38
wikipedia_all_enga	17,421	35,165	36,760	53.93	61.11
	723,612	7,580,187	7,978,748		

IRIS – Wish List (Future Work)

- getting feedback on the system
- publications on IRIS
- improving translation quality
  → using more parallel/monolingual Irish data
- improving translation time
  → ignoring translation candidates with low probability, direct recasing, ...
- evaluation of translation quality
  → manual evaluation of machine translation

http://server1.nlp.insight-centre.org/tetra/

IRIS - English-Irish Translation System

Mihael Arcan, UNLP, Insight@NUI Galway
mihael.arcan@insight-centre.org