

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

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/>

Research 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

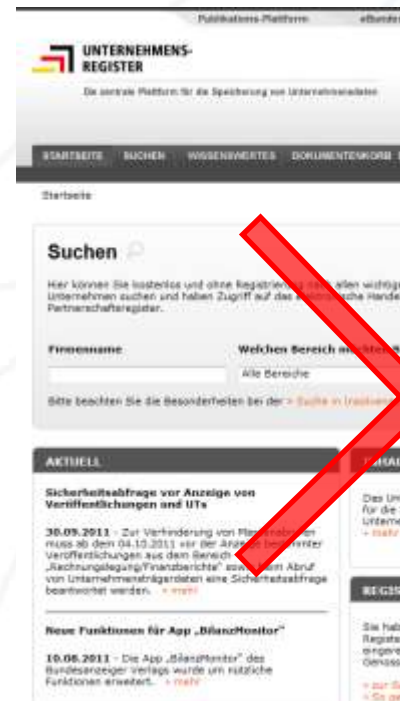
<http://nlp.insight-centre.org/>

Motivation of my Study

Business Information query in English
equity@en



The screenshot shows the NYSE Euronext website interface. At the top, there are navigation tabs for 'Home', 'Listings', 'Equity', 'Bonds', 'Futures/Options', 'Market Information', 'Regulation', and 'Business Services'. Below this, there's a section for 'Market indicators' with a line chart showing data for 'FTSE World 50'. To the left, there's a search bar with 'intrasense' entered. Below the search bar, there's a table of market indicators with columns for 'Index', 'Last', and '%'. The table includes entries like 'FTSE World 50', 'NASDAQ 100', 'DAX', 'S&P 500', 'EURO STOXX 50', 'CAC 40', and 'FTSE 100'. There are also sections for 'My Portfolio' and 'My Watchlist'.

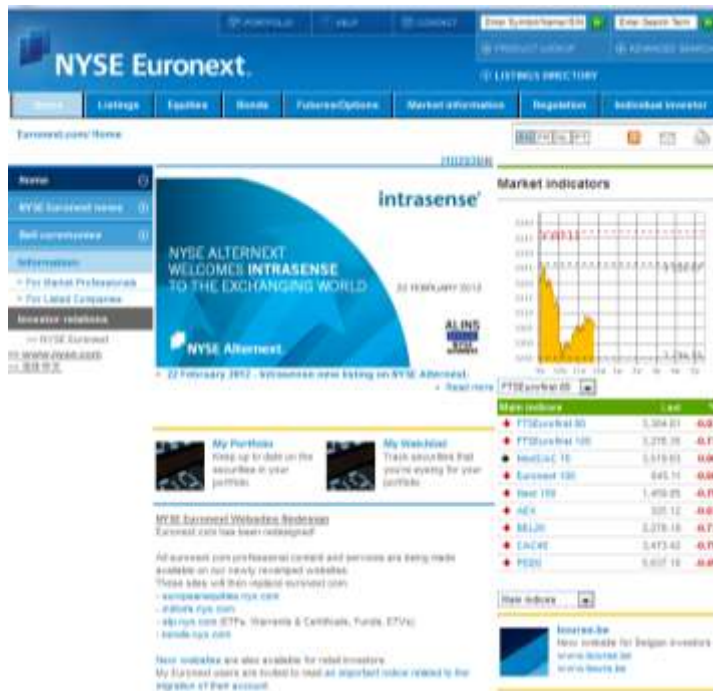


The screenshot shows the 'UNTERNEHMENSREGISTER' website. It features a search bar with the text 'Suchen' and a large red 'X' overlaid on the page. Below the search bar, there's a section for 'Aktuelle' (Current) with a list of news items. The first item is dated '30.09.2011' and discusses 'Sicherheitsabfrage vor Anzeige von Veröffentlichungen und UTs'. The second item is dated '10.06.2011' and discusses 'Neue Funktionen für App „BilanzMonitor“'. The website also has a navigation menu at the top with 'Startseite', 'Suchen', 'Wissenswertes', and 'Dokumentations-PP'.



Motivation of my Study

Business Information query in English
equity@en -> Google Translate -> Gerechtigkeit@de



The screenshot shows the NYSE Euronext website interface. At the top, there's a navigation bar with 'NYSE Euronext' and various menu items like 'Listings', 'Equities', 'Bonds', 'Futures/Options', 'Market Information', 'Regulation', and 'Business Services'. Below this, there's a main content area with a large banner for 'NYSE ALIANT WELCOMES INTRASENSE TO THE EXCHANGING WORLD'. To the right, there are 'Market Indicators' with a line chart showing price fluctuations. Below the chart is a table of market indicators with columns for 'Market Indicators', 'Last', and '%'. The table lists various indices like FTSEurofirst 50, FTSEurofirst 100, Next 100, etc. There are also sections for 'My Portfolio' and 'NYSE Euronext Webinars'.

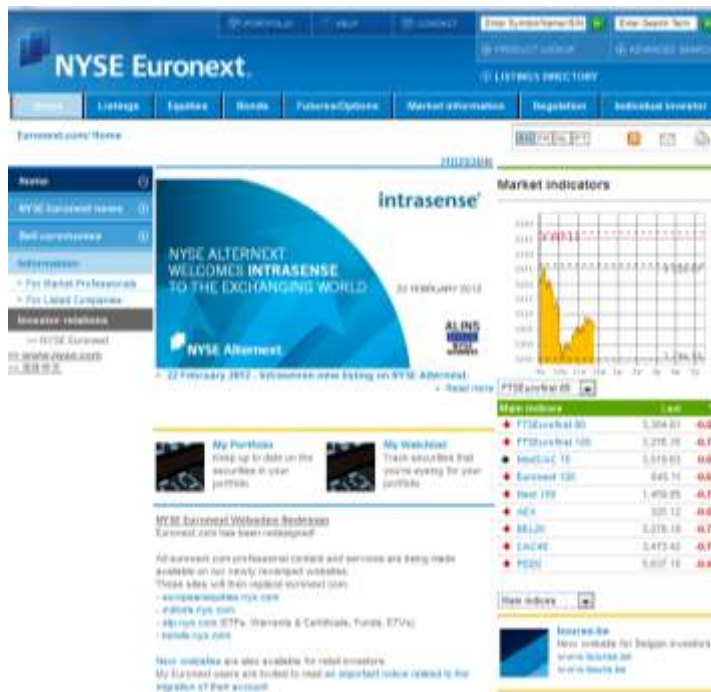


The screenshot shows the 'UNTERNEHMENS-REGISTER' website. The header includes the logo and the text 'Die zentrale Plattform für die Speicherung von Unternehmensdaten'. Below the header, there's a search section titled 'Suchen' with a search bar and a 'Suchen' button. The search results area contains a list of news items under the heading 'AKTUELL'. The first item is 'Sicherheitsabfrage vor Anzeige von Veröffentlichungen und UTs' dated 30.09.2011. The second item is 'Neue Funktionen für App „BilanzMonitor“' dated 10.06.2011. The website also features a 'REGISTRIERT' section on the right side.



Motivation of my Study

Business Information query in English
equity@en -> domain aware SMT-> Eigenkapital@de



The screenshot shows the NYSE Euronext website interface. At the top, there's a navigation bar with 'NYSE Euronext' and various menu items like 'Listings', 'Equities', 'Bonds', etc. Below the navigation, there's a main content area featuring a large advertisement for 'intrasense' with the text 'NYSE ALTERNEXT WELCOMES INTRASENSE TO THE EXCHANGING WORLD'. To the right of the ad, there are 'Market indicators' with a line chart showing data over time. Below the chart, there's a table of market indicators with columns for 'Name', 'Last', and '%'. The table lists various indices like FTSEurofirst 50, FTSEurofirst 100, Nextec 10, Eurostoxx 500, Next 100, AEX, BEL20, CAC40, and PSI20. On the left side, there are sections for 'My Portfolio' and 'My Watchlist'. At the bottom, there's a section for 'NYSE Euronext Webinars' and 'New indices'.



The screenshot shows the 'Unternehmensregister' website. At the top, there's a navigation bar with 'Unternehmensregister' and 'Suchen'. Below the navigation, there's a search section with the text 'Hier können Sie kostenlos und ohne Registrierung nach allen wichtigen Unternehmen suchen und haben Zugriff auf das elektronische Handels-, Partnerschaftsregister.' There are input fields for 'Firmenname' and 'Welchen Bereich möchten Sie' with a dropdown menu set to 'Alle Bereiche'. Below the search section, there's a 'AKTUELL' section with news items. The first news item is 'Sicherheitsabfrage vor Anzeige von Veröffentlichungen und UTs' dated 30.09.2011. The second news item is 'Neue Funktionen für App „BilanzMonitor“' dated 10.06.2011. On the right side, there's a 'INHALT' section with links to 'Das Unter-für die So-Unternehm...' and 'REGISTRI'.



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

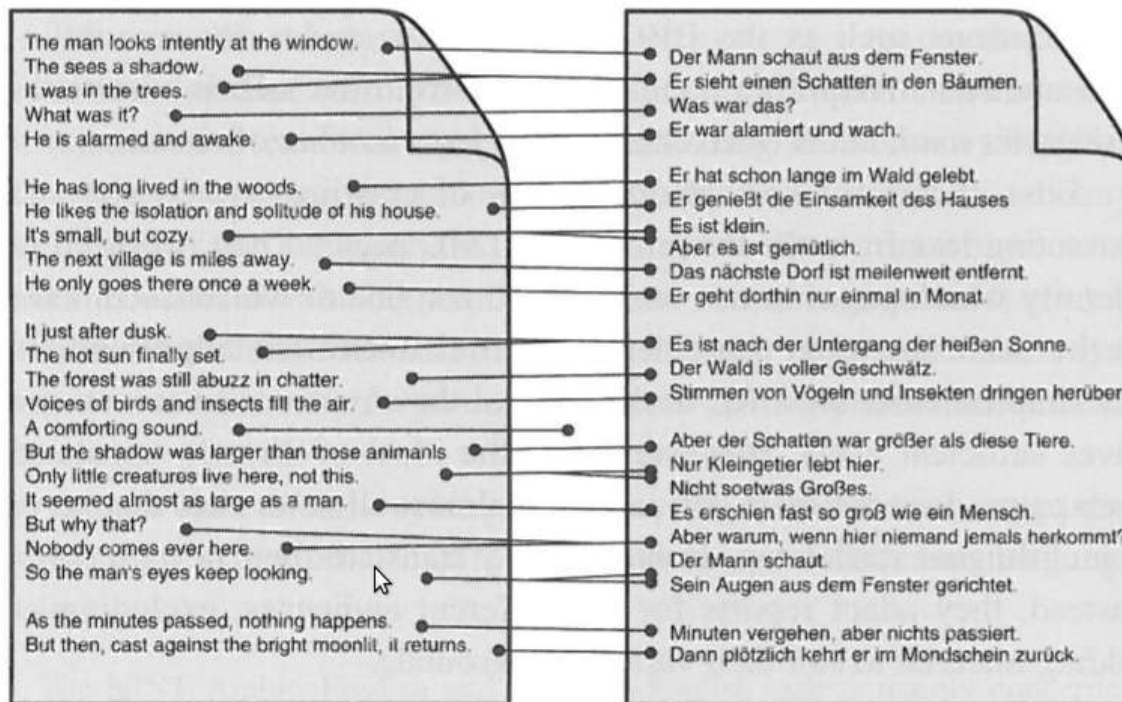
- takes care of fluency (and lexical choice) in the target language

-4.868038	anlagevermögen	-0.1317768
-----------	----------------	------------

Training Data (Parallel Corpora) for SMT

Source document

Target document



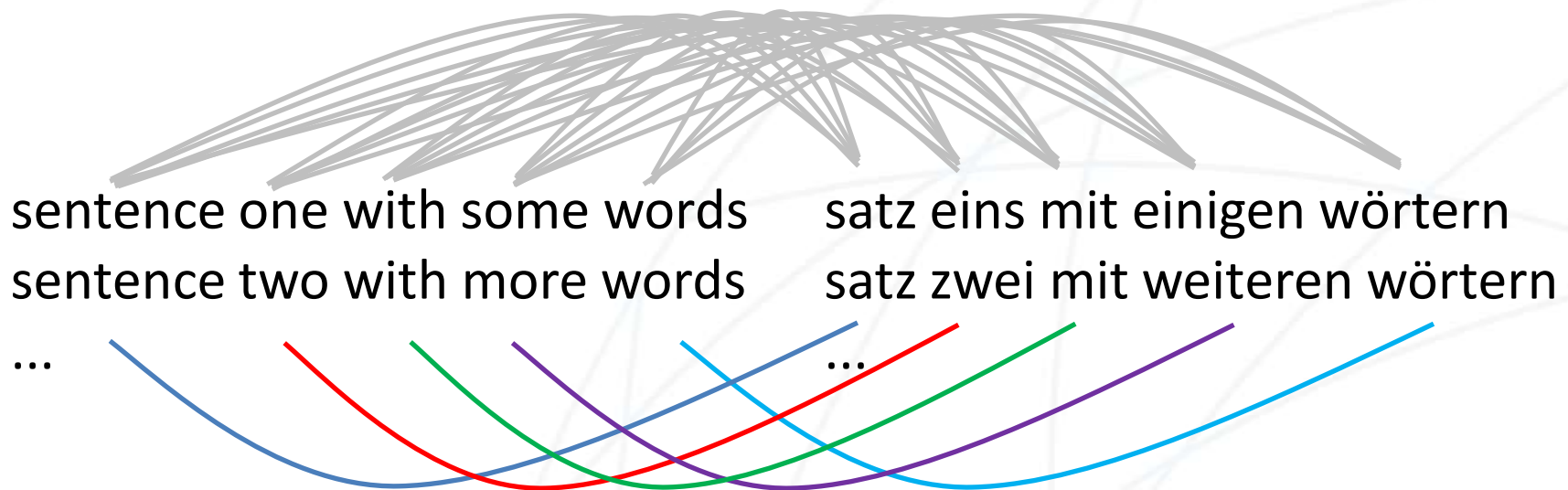
Training Data for SMT

- Sentence aligned parallel data

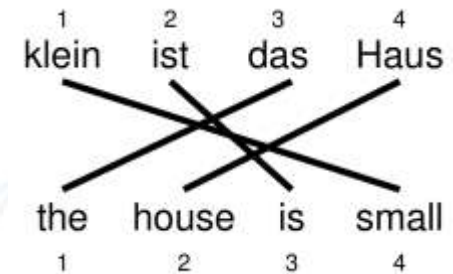
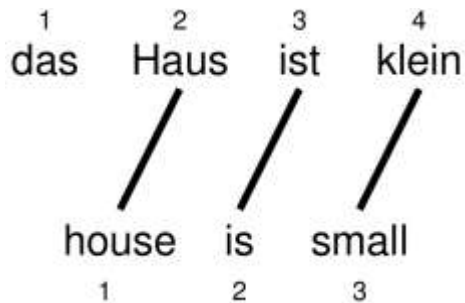
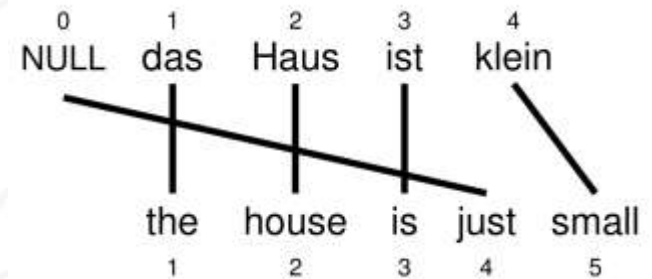
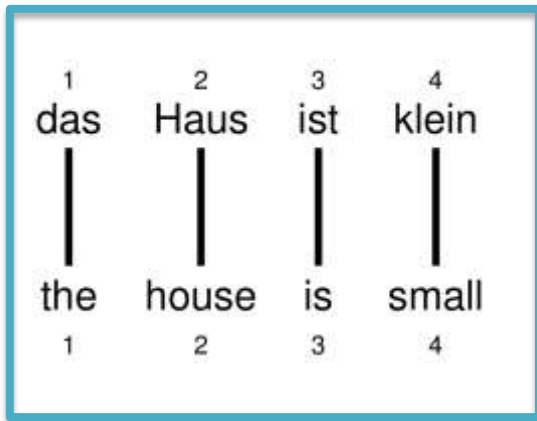


Training Data for SMT

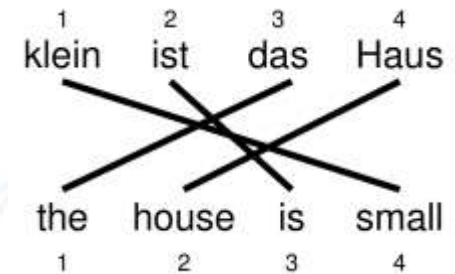
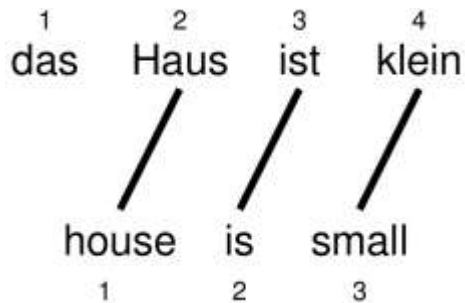
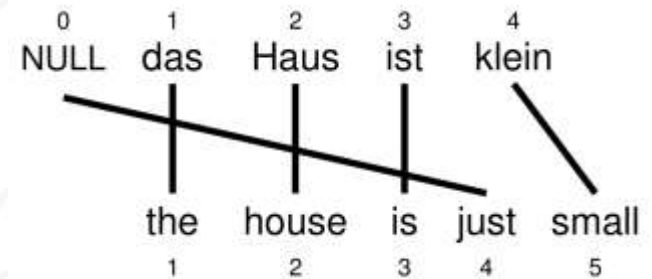
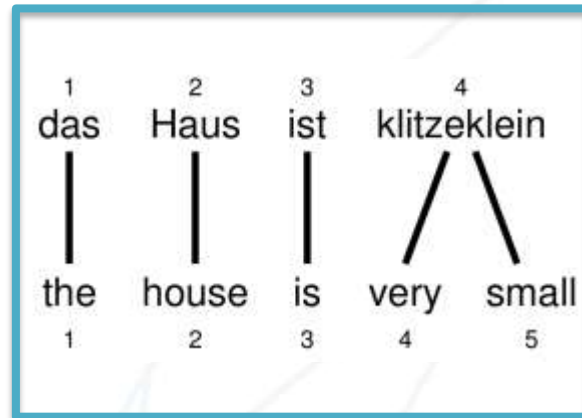
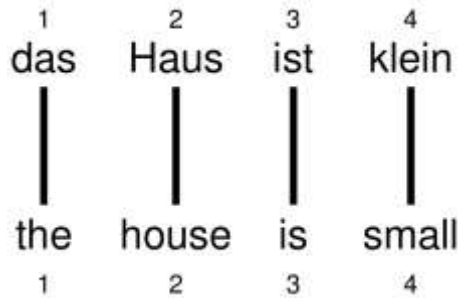
- Sentence aligned parallel data → Word Alignment



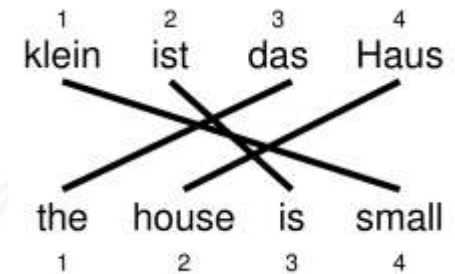
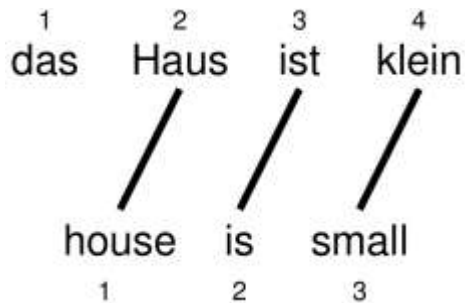
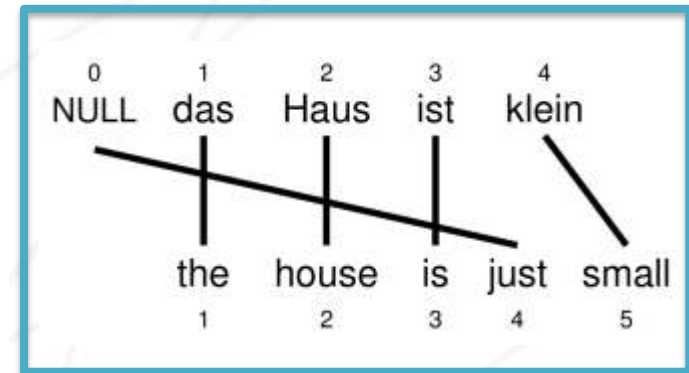
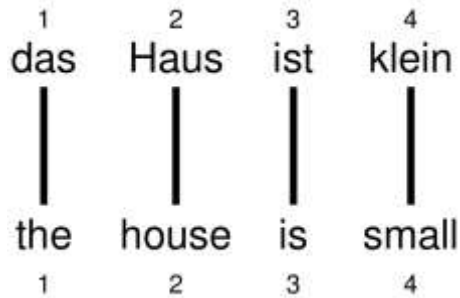
Word Alignment Scenarios



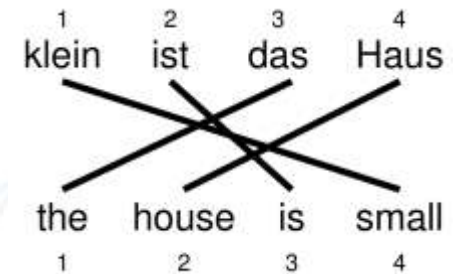
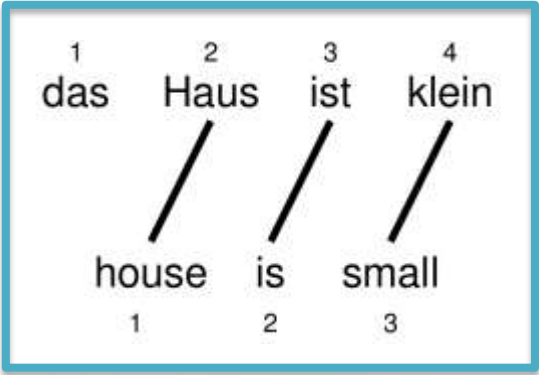
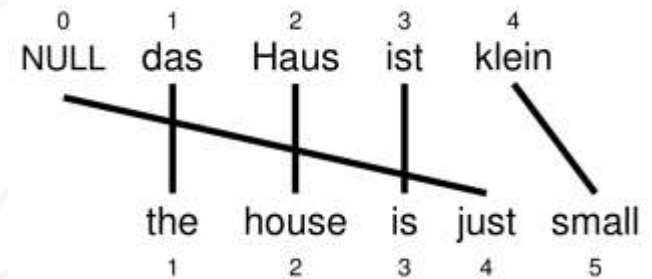
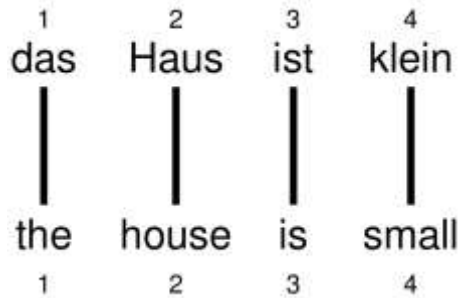
Word Alignment Scenarios



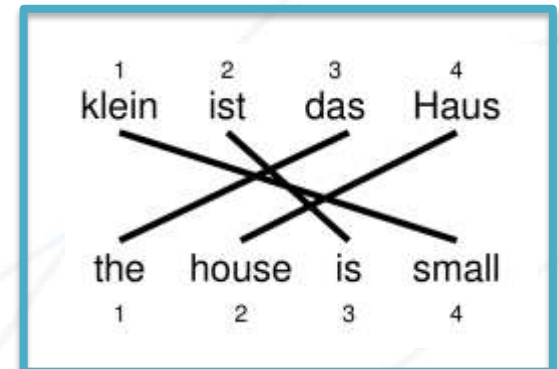
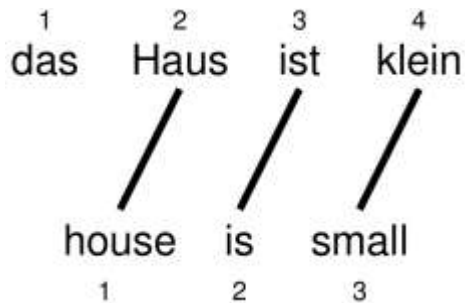
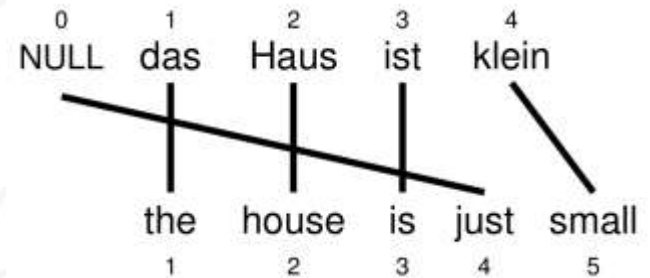
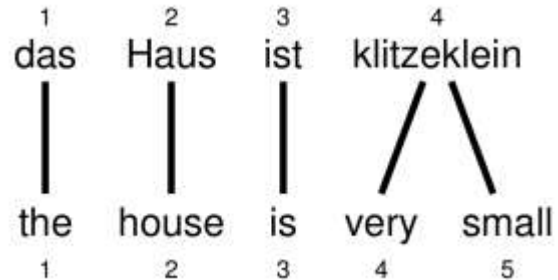
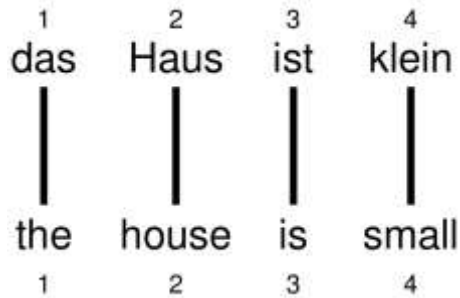
Word Alignment Scenarios



Word Alignment Scenarios



Word Alignment Scenarios



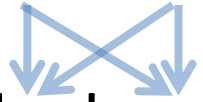
Word Alignment with IBM Models


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

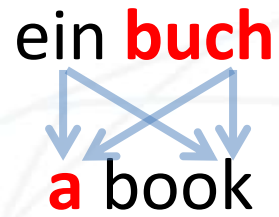
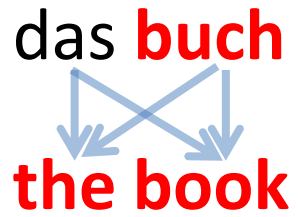
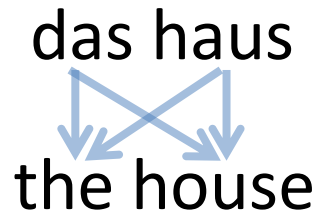
Training data:

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

das haus

the house

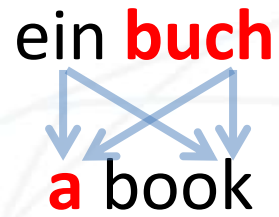
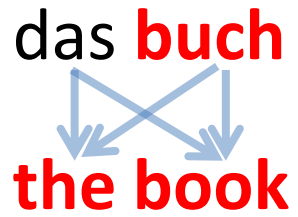
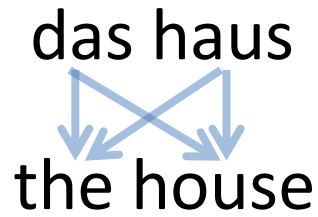
das **buch**

the book

ein **buch**

a book



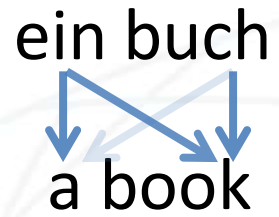
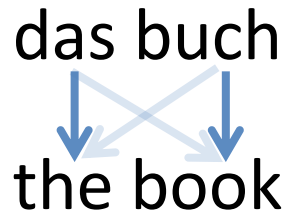
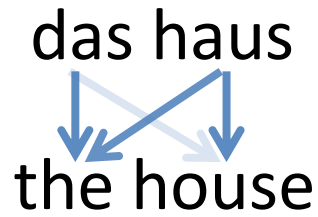
e	f	Initial	1st iter.	2nd iter.	3rd iter.	...	Final
the	das	0.25					1
book	das	0.25					0
house	das	0.25					0
the	buch	0.25					0
book	buch	0.25					1
a	buch	0.25					0
book	ein	0.25					0
a	ein	0.25					1
the	haus	0.25					0
house	haus	0.25					1

Learning Phase



e	f	Initial	1st iter.	2nd iter.	3rd iter.	...	Final
the	das	0.25					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



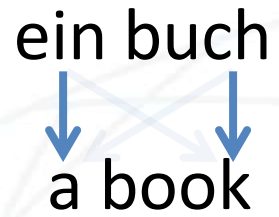
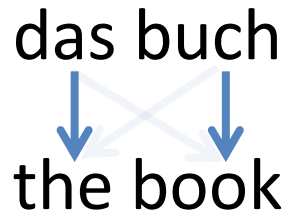
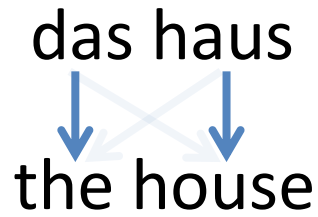
e	f	Initial	1st iter.	2nd iter.	3rd iter.	...	Final
the	das	0.25	0.50				1
book	das	0.25	0.25				0
house	das	0.25	0.25				0
the	buch	0.25	0.25				0
book	buch	0.25	0.50				1
a	buch	0.25	0.25				0
book	ein	0.25	0.50				0
a	ein	0.25	0.50				1
the	haus	0.25	0.50				0
house	haus	0.25	0.50				1

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			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

das haus
↓ ↓
the house

das buch
↓ ↓
the book

ein buch
↓ ↓
a book

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

das haus
↓ ↓
the house

das buch
↓ ↓
the book

ein buch
↓ ↓
a book

Decoding (translating) using the lexical probabilities:

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

das haus
↓ ↓
the house

das buch
↓ ↓
the book

ein buch
↓ ↓
a book

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

schlechte⁶ bank²

bad⁶ bench⁴ 0.1239

bad⁶ bank² 0.1239



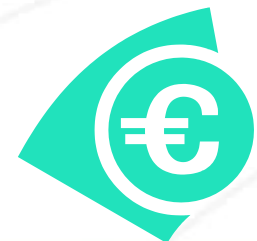
Generic Models in SMT

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
<u>grüne</u> ⁷ <u>bank</u> ⁴	<u>green</u> ⁷ <u>bench</u> ⁴

schlechte⁶ bank²

bad⁶ bench⁴ 0.1918 ✘

bad⁶ bank² 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

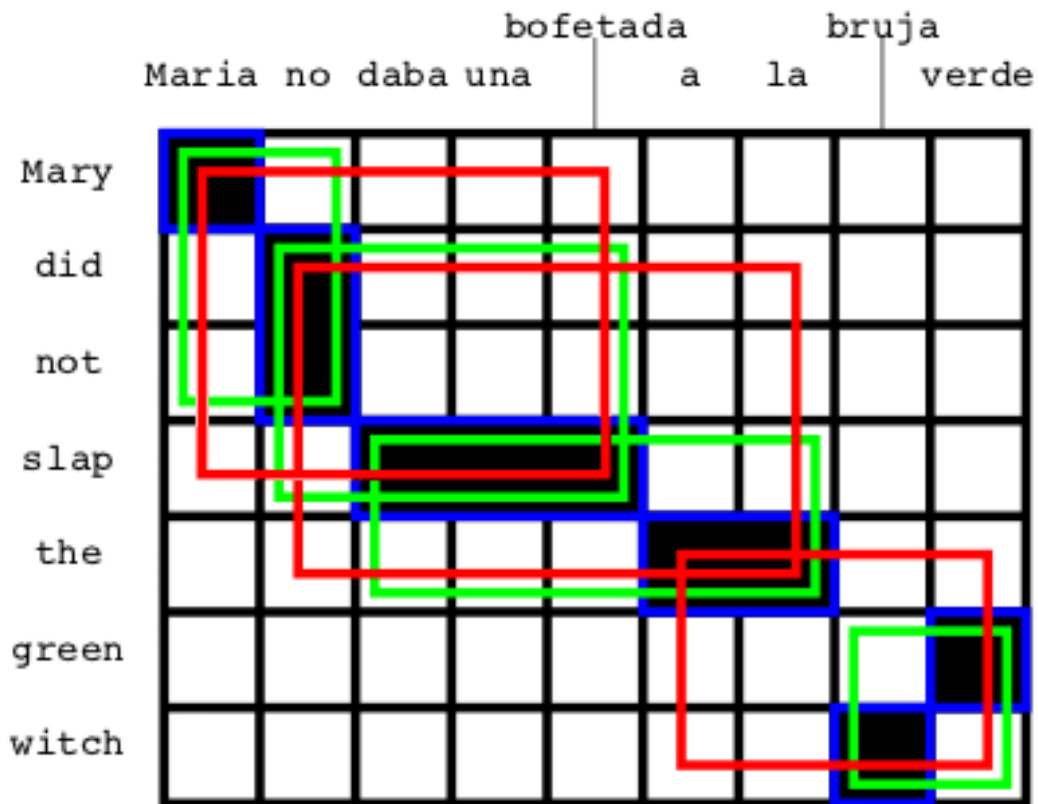
schlechte⁶ bank²

bad⁶ bank² 0.1918

bad⁶ bench⁴ 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

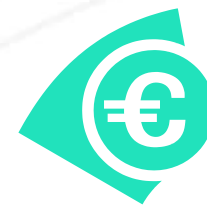
freundliche¹ bank²

friendly¹ bench⁴

0.1576 

friendly¹ bank²

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_bank¹

friendly_bank¹

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

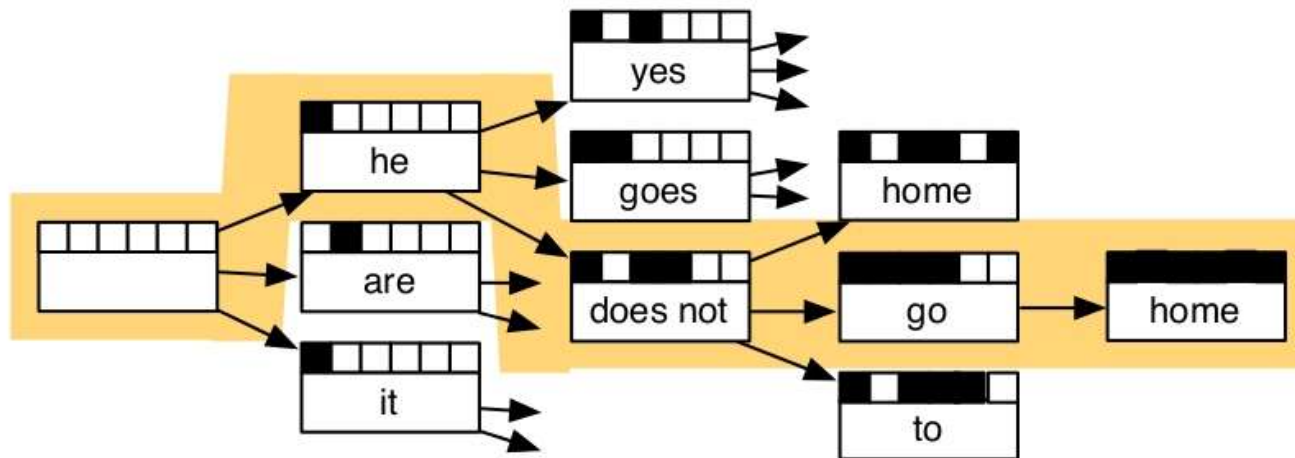
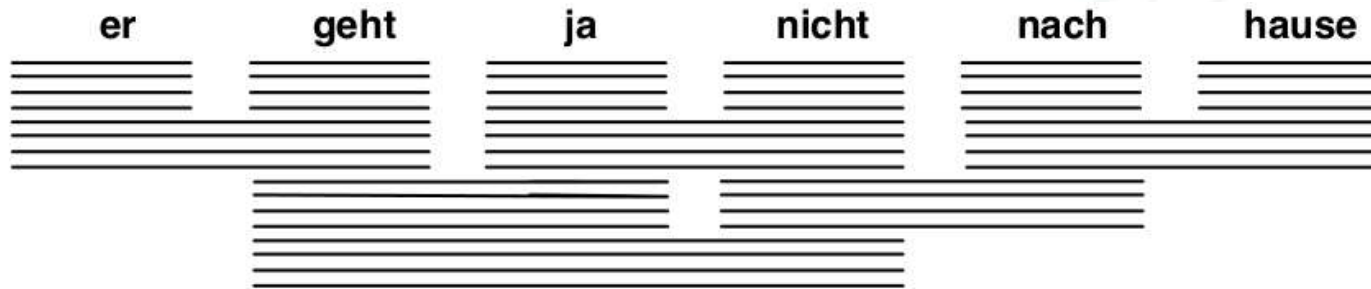
- das

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*



The screenshot shows the IRIS English-Irish Translation System interface. At the top, there is a navigation bar with the IRIS logo and the text "English-Irish Translation System". The navigation bar includes links for Home, Irish, About, Team, and Other Projects. The main content area is divided into two columns. The left column contains a text input area with the instruction "Add English or Irish text to be translated, or use the text examples:". Below this are two buttons: "English Sentence" and "Irish Sentence". A large empty text box is provided for input. At the bottom of this column is a "Translate" button. The right column contains a text block explaining the system's purpose: "The IRIS Translation System was developed to translate text from English into Irish and vice versa. Additionally, IRIS allows to upload English and/or Irish monolingual or parallel text to update the translation models. For uploading a new English/Irish dataset, [click here to open a new window.](#)". Below this text are two bar charts. The top chart is titled "English-Irish SMT system Evaluation" and shows BLEU scores for 9 datapoints of new added data. The bottom chart is titled "Irish-English SMT system Evaluation" and shows BLEU scores for 9 datapoints of new added data. Both charts show scores consistently above 0.6. At the bottom of the interface, there are logos for Insight, NUI Galway OE Gaillimh, and SFI. A footer note states: "This resource has been funded by Grant No. SFI/12/RC/2289 - INSIGHT - National University of Ireland, Galway".

<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)
- allows 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