

Statistical (& Neural) Machine Translation

Prof. Andy Way

ADAPT Centre,
School of Computing,
Dublin City University,
Dublin 9, Ireland

andy.way@adaptcentre.ie

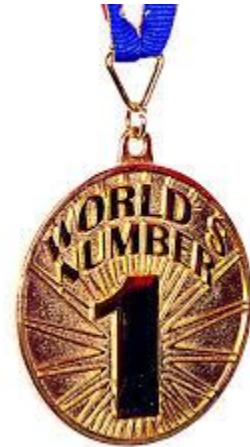
<http://www.computing.dcu.ie/~away/>

Overview

- Why MT?
- How are people using MT?
- What's the Role of the Human Translator?
- Why Corpus-Based MT?
- How might you go about translating languages you have no knowledge of?
- The Importance of Data

DCU are *the world leaders* in MT!

- MT thought-leaders
- Large amounts of funding for MT
- World-leading MT performance:
 - CWMT-2018: **2x 1st** place
 - WMT-2017: **3x 1st** place
 - WMT-2016: **3x 1st** place
 - WMT-2015: **3x 1st** place
 - WMT-2014: **14x 1st** places
- Leading performance in tweet, image, and dialogue translation!



Why we're the best at what we do ...

“Of the international leading MT research groups, ADAPT is the one with the closest connection to deployment of the technology. ADAPT hosts some of the leading researchers on human-computer interaction in MT, and the integration of MT in the workflow of professional translators.”

“Research on machine translation is very broad, including problems such as multi-modal translation, dialogue translation, etc.”

“Researchers in the ADAPT centre, building on earlier efforts in the CNGL centre, have two unique strengths:

- the commercialization of MT including tackling various deployment and usage problems,*
- considering human factors in translation.”*

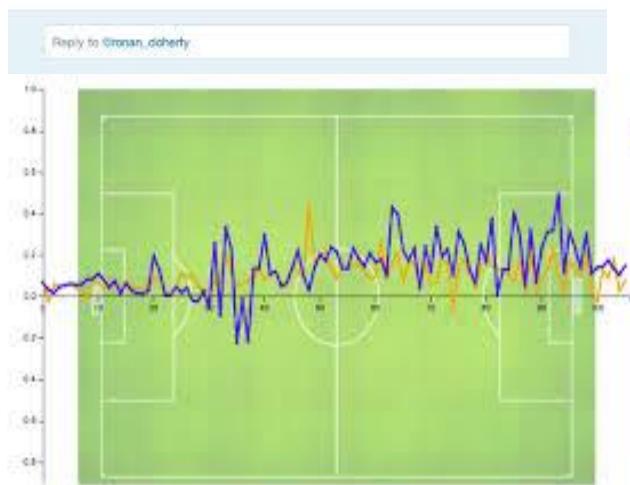
Industrial & Societal Impact



Ronan Doherty
@ronan_doherty

Comhghairdeas Deutschland! Cluiche a chuir ceann ceart ar an chraobh.
#CornAnDomhain #WordCup2014

Reply · Retweet · Fav · More · 2:47 PM - 13 Jul 2014



Irish Independent THE IRISH TIMES



silicon republic®

Volume

The 2014 FIFA World Cup was the biggest event yet for Twitter with **672 million tweets**

Requested translation from Twitter (words)	 	 	 	Grand Total from all World Cup matches
	6,459,830	5,141,360	4,847,590	85,047,110

Top 3 languages

English

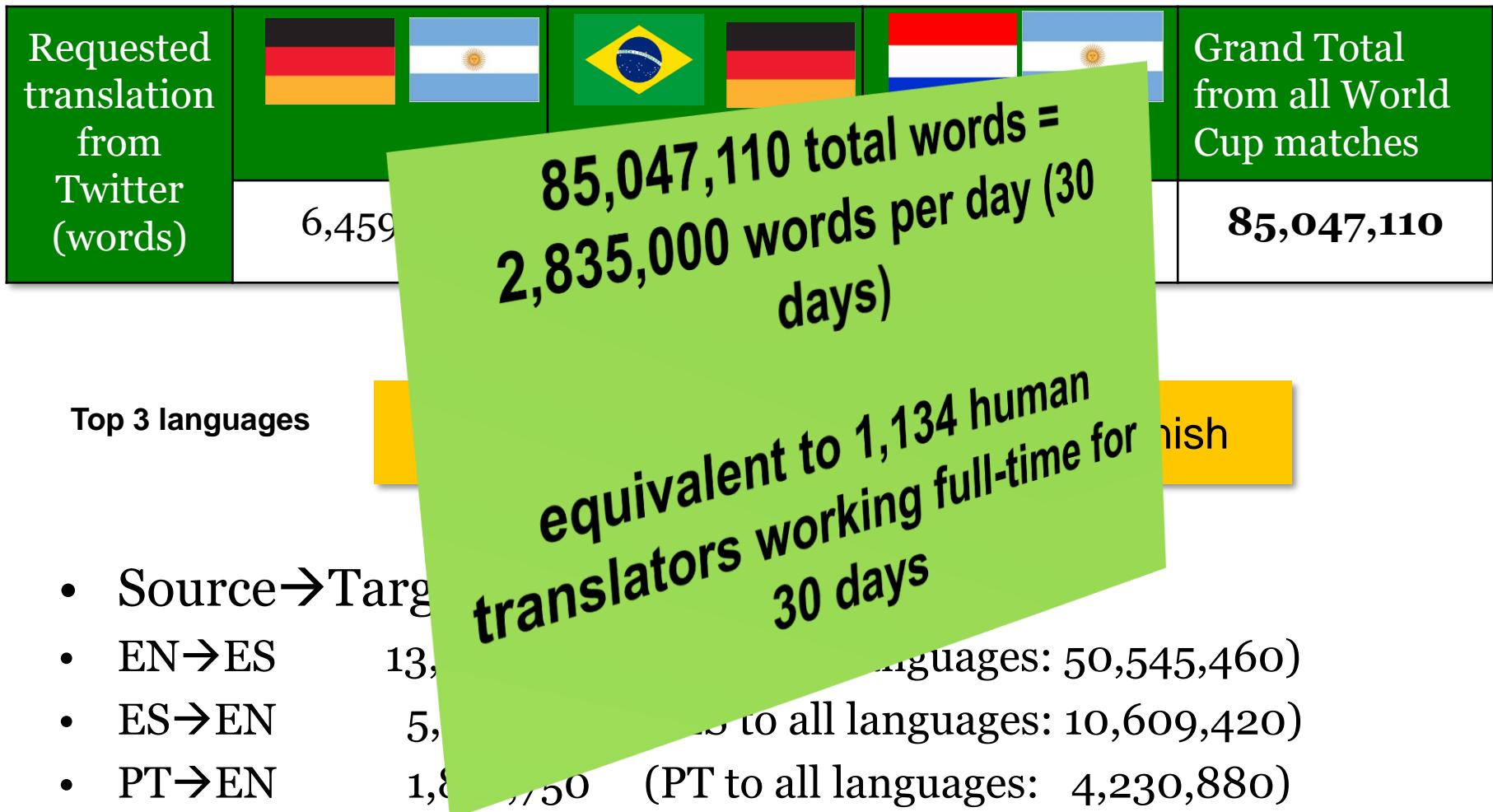
Portuguese

Spanish

- Source→Target traffic:
- EN→ES 13,614,450 (EN to all languages: 50,545,460)
- ES→EN 5,569,200 (ES to all languages: 10,609,420)
- PT→EN 1,831,750 (PT to all languages: 4,230,880)

Volume

The 2014 FIFA World Cup was the biggest event yet for Twitter with **672 million tweets**



13 languages and 24 language pairs



24 MT systems trained on Microsoft Translator Hub and used by



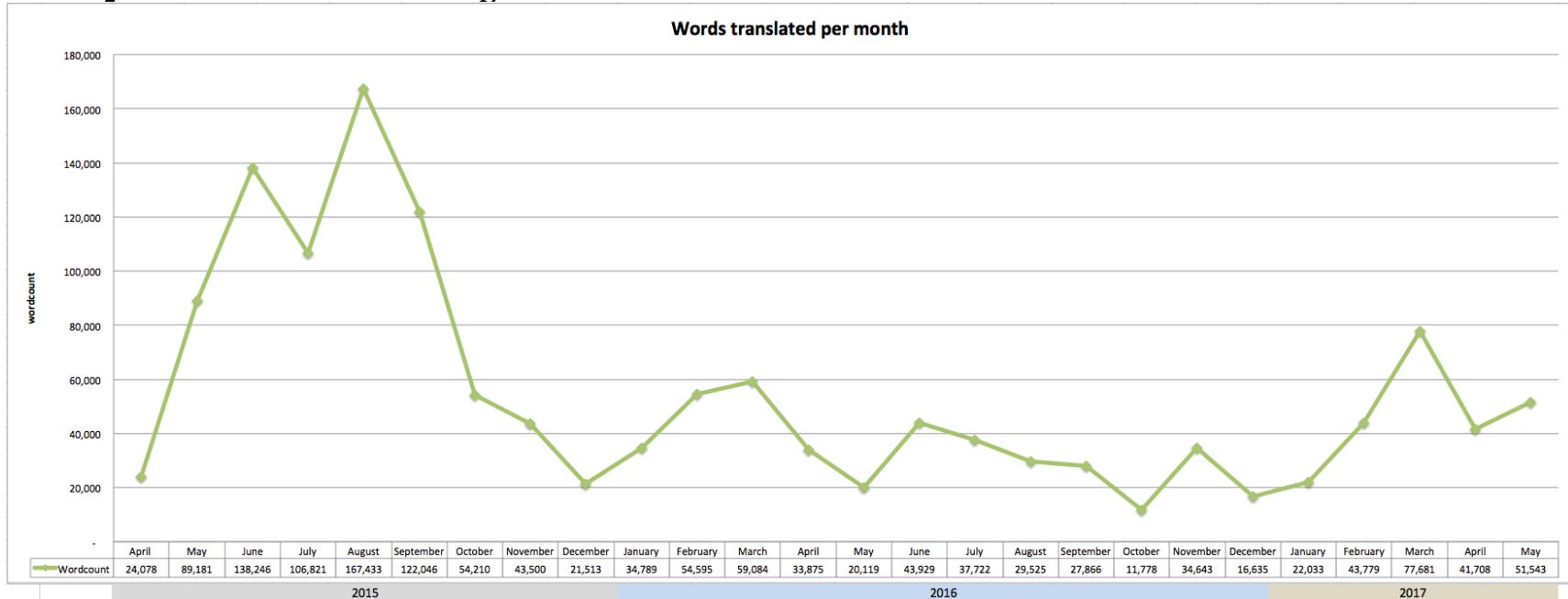
FurHat demo for hotel booking



Societal Impact



Tapadóir MT System



- Internal trials in DAHG indicated output was useable → full roll-out
- Productivity increasing; no detrimental effect on **any** translators studied
- Integrated seamlessly with existing translation workflow: **1.5 million words** so far



SIZE MACHINE TRANSLATION MARKET IS \$250 MILLION – TAUS PUBLISHES NEW MARKET REPORT

August 26, 2014, Amsterdam - TAUS estimates the size of the machine translation market at \$250 million in its newly published machine translation market report. The 60 page report offers a detailed overview of all facets of the machine translation with sections on the different types of offerings, the players, open-source systems, challenges, opportunities and trends.

"The size of the MT market is relatively small compared to its innovation power and impact", says Jaap van der Meel, one of the co-authors of the report. *"MT technology is a key enabler and a force multiplier for new services and growth. MT technology finds a high adoption rate among language service providers. Innovative companies in information technology and other sectors are converging MT technology in new applications and products or they use MT to enhance their existing products."*

For this market report TAUS has identified 65 different MT operators. More than 80 companies responded to the surveys and the TAUS team interviewed 37 users and developers of MT. The largest MT providers in alphabetical order are: [CSLI](#), [Google](#), [IBM](#), [LionBridge](#), [Microsoft](#), [PROMT](#), [Raytheon BBN](#), [SDL](#), [Smart Communications](#), [SYSTRAN](#). The MT market is a vibrant sector with new companies entering the market place and long-term players being acquired. Around 20 of the 65 identified MT players started business in the last five years. At the start of this year [SYSTRAN](#) was acquired by [CSLI](#) from South Korea and [AppTek](#) was acquired by [eBay](#).

"The dynamics in the MT market have changed dramatically in the last five years", says Achim Ruopp, product development manager at TAUS and co-author of the report. *"The increased availability of easy to use and integrate MT with sufficient quality has ignited the emergence of new business models. This has been promoted by many new MT suppliers that base their offering on the open source statistical MT system [Moses](#). A bigger impact still has come from some of the internet giants like Google, Microsoft and Yandex that offer free or very cheap MT services."*

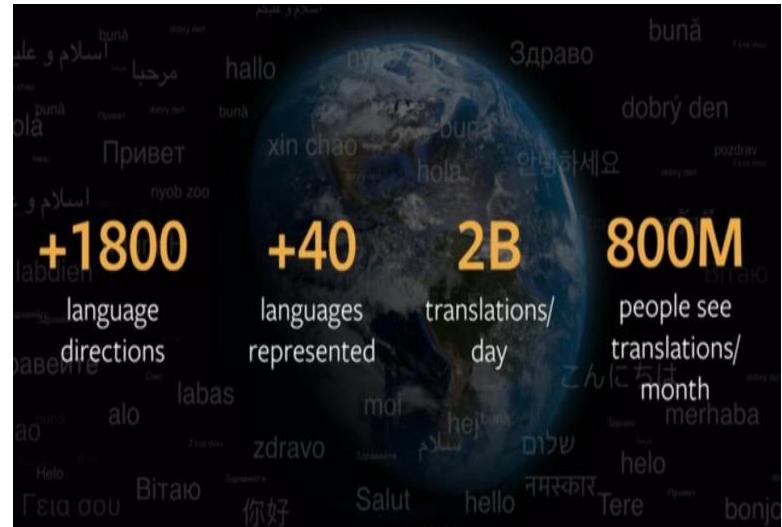
The [TAUS Machine Translation Market Report](#) is available on the TAUS website.



Highlights from TAUS Claims

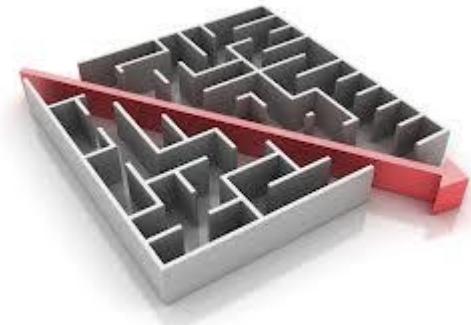
- Size of the market: €250,000,000
- “MT technology is a **key enabler** and a **force multiplier** for **new services**”
- “**Innovative companies** in IT and other sectors are converging MT technology in new applications and products or they **use MT to enhance their existing products**”
- “The increased availability of **easy to use** and **integrate MT with sufficient quality** has ignited the emergence of new business models. This has been promoted by many **new MT suppliers** that base their offering on the open source statistical MT system **Moses**”

MT is being used every day ...



- . Google Translate provides a billion translations a day for 200 million users
- . Amount of text translated daily is more than what's in a million books
- . Surpasses what professional translators handle in a year

Client-customised engines



- Improve productivity,
- Translate content previously not feasible due to time or cost constraints,
- Reduce time to market, and
- Reduce translation costs.

Lots of successful case studies



- Adobe & ProMT
- Church of Jesus Christ of Latter-day Saints & Microsoft Translator Hub
- Dell & Safaba/we localize
- DuDu & CapitaTI
- Ford & Systran/SAIC
- Sajan & Asia Online
- text&form & LucySoft

The time for MT is



Not everyone agrees ...



Why Corpus-Based MT?

- the (relative) failure of rule-based approaches
- the increasing availability of machine-readable text
- the increase in capability of hardware (CPU, memory, disk space) with decrease in cost

Why is MT Hard?

- Human languages are:
 - Elegant
 - Efficient
 - Flexible
 - Complex
- One word/sentence may mean many things
- Many ways of saying the same thing
- Meaning depends on context
- Literal and figurative language (metaphor)
- Language and culture (different ways of conceptualising the same thing)
- Word order
- Morphology
- ...

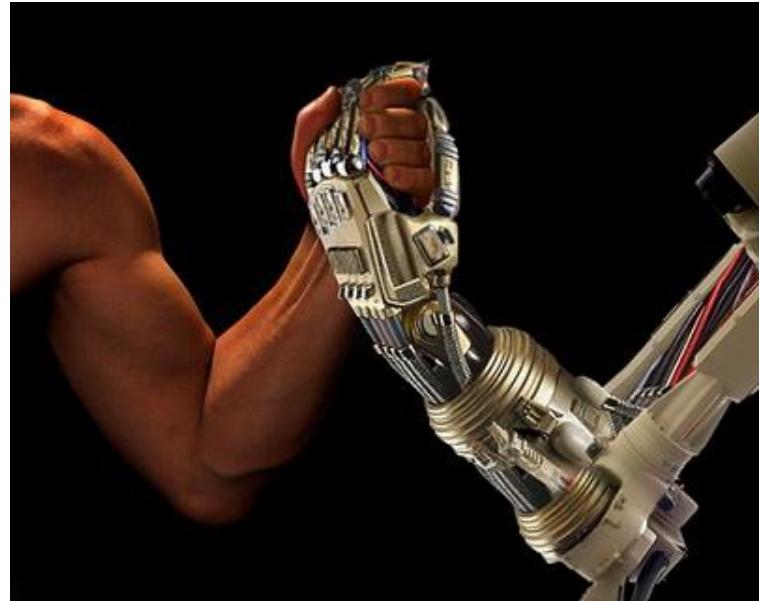


Image: <http://workingtropes.lmc.gatech.edu/wiki/index.php/File:Man-vs-machine.jpg>
License: CC BY-NC-SA 3.0

Why is MT Hard?

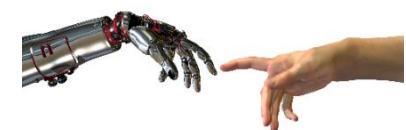
Newspaper Headlines:

1. Minister Accused Of Having 8 Wives In Jail
2. Juvenile Court to Try Shooting Defendant
3. Teacher Strikes Idle Kids
4. Miners refuse to work after death
5. Local High School Dropouts Cut in Half
6. Red Tape Holds Up New Bridges
7. Clinton Wins on Budget, but More Lies Ahead
8. Hospitals Are Sued by 7 Foot Doctors
9. Police: Crack Found in Man's Buttocks

Thanks to Chris Manning

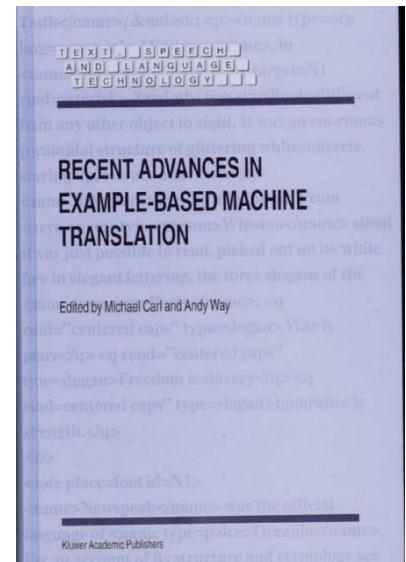
Language & Translation is Complex

- Language/translation is complex
- We cannot compute it exactly
- We tried: rule-based MT and LT ...
- What do we do *now*?
- Machine Learning
 - Learns from **data** \Rightarrow data is (mostly) all important
 - Approximate solution \Rightarrow not perfect, needs help
 - Human Professional Translators
 - Post-editing
 - Automated Translation \neq Automatic



Types of Corpus-based MT

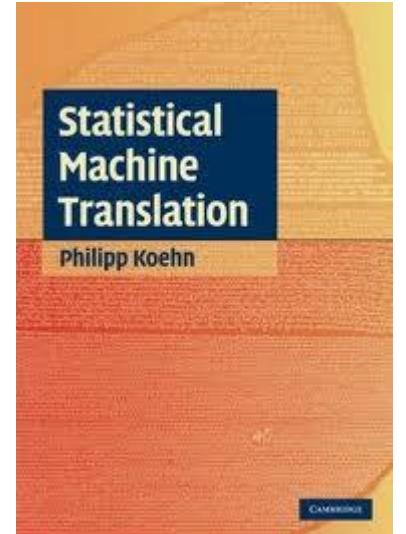
Example-Based MT (Nagao, 1984)



Statistical MT

- 1988: word-based (IBM)
- 2002—now: phrase-based (Moses)
- 2005—now: tree-based (Hiero)

Neural MT (1997 ... and 2013 onwards)



Prerequisite

A prerequisite for Data-Driven MT (and also TM, which is *not* MT, but rather CAT):

- Example-Based MT (EBMT)
- Statistical MT (SMT) & Neural MT (NMT)
- Hybrid Models which use some probabilistic processing

is a *parallel corpus* (or *bitext*) of aligned sentences.

Parallel data prerequisite for corpus-based MT

PROMT Translation Memory Manager - [Business - entire TM database]

IM Database Edit Tools View Window Help

Translation Memory Database Business - entire TM database

Business
Idioms
English-Spanish
Idioms
French-English
Idioms
German-English
Business
Idioms
Travel
Portuguese-English

TM Database Properties

Title

Name	Business
Source	German
Translation	English
Comments	

Statistics

Find Text

Find:
Replace with:
 Match case Field for search:
 Regular expressions Search direction:

Find Replace Replace All Help

Ready German-English Record 1:3266

The screenshot shows the PROMT Translation Memory Manager application window. On the left, there's a sidebar with a tree view of available translation memory databases: Business, Idioms, English-Spanish, French-English, German-English, Portuguese-English, and Travel. Below this is a 'TM Database Properties' section with fields for Title, Name (Business), Source (German), Translation (English), and Comments. At the bottom of the sidebar is a 'Statistics' section which is currently empty. The main area displays a table titled 'Business - entire TM database' with two columns: 'Source Text' and 'Translated Text'. There are four entries in the table:

Source Text	Translated Text
Wir wären Ihnen sehr dankbar, wenn Sie diesen Zahlplan annehmen würden.	We should be very grateful to you for accepting this payment plan.
Wir sind über den Ton Ihres Schreibens sehr ungehalten.	We are feeling extremely indignant at the tone of your letter.
Wir können Ihnen keinen weiteren Aufschub gewähren.	We will not be able to allow you any further delay.
Es lief wie am Schnürchen.	It went like clockwork.

At the bottom of the main window, a 'Find Text' dialog is open. It contains fields for 'Find' and 'Replace with', checkboxes for 'Match case' and 'Regular expressions', dropdowns for 'Field for search' (set to 'Source Text') and 'Search direction' (set to 'Down'), and buttons for 'Find', 'Replace', 'Replace All', and 'Help'.

Parallel data prerequisite for corpus-based MT

PROMT Translation Memory Manager - [Business - entire TM database]

IM Database Edit Tools View Window Help

Translation Memory Database Properties

Business
Idioms
English-Spanish
Idioms
French-English
Idioms
German-English
Business
Idioms
Travel
Portuguese-English

TM Database Properties

Title

Name	Business
Source	German
Translation	English
Comments	

Statistics

Find

Find:
Replace:
 Match case
 Regular expressions

Field for search:

Search direction:

Replace

Replace All

Help

very grateful to you for payment plan.

extremely indignant at the tone

able to allow you any further

ickwork.

I ❤️ Translators

Ready

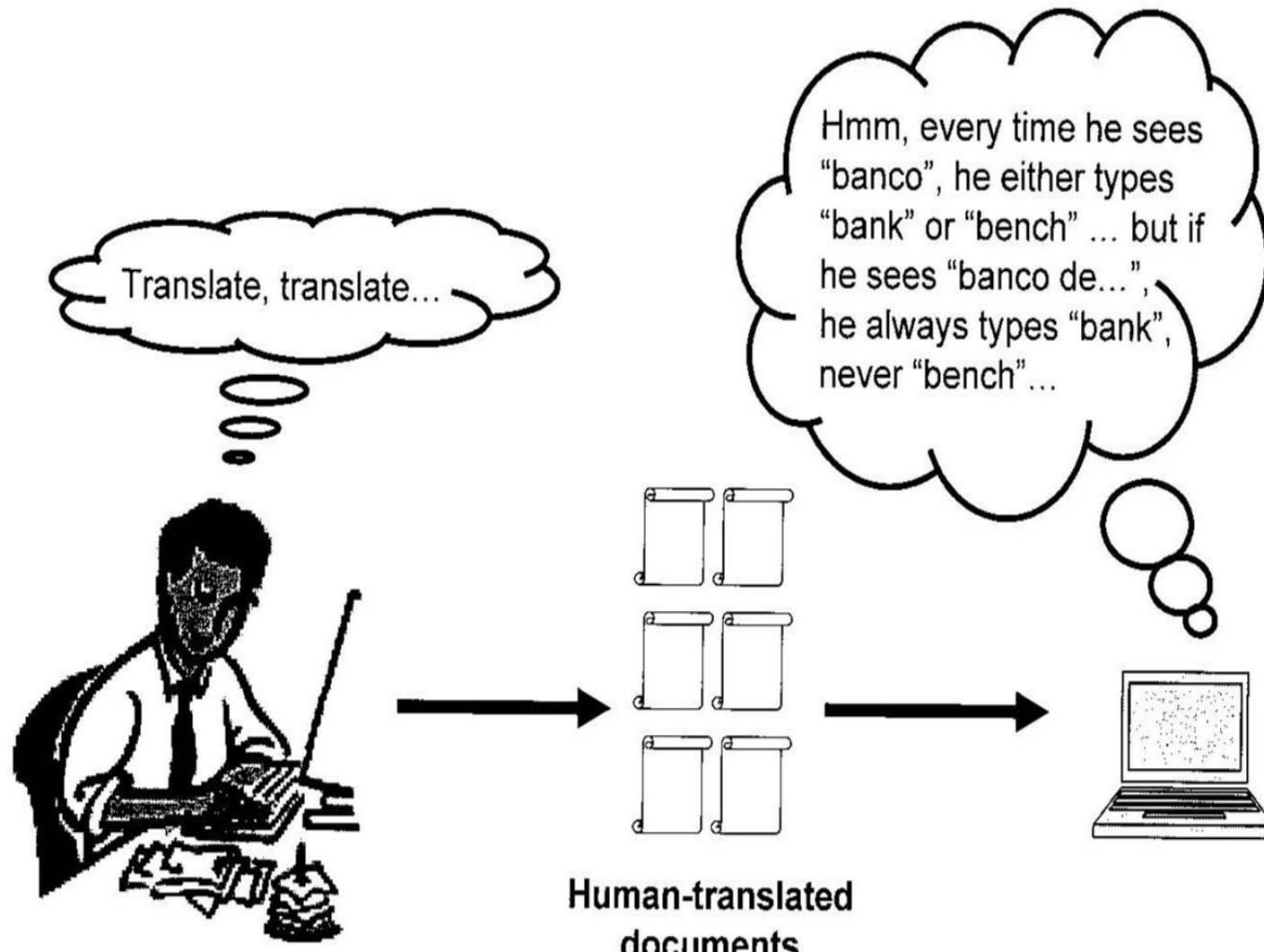
German-English Record 1:3266

So how does SMT work?

How might *you* go about translating between two languages you know

nothing about?!

Statistical Machine Translation



Thanks to Kevin Knight ...

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: farok crrrok hihok yorok clok kantok ok-yurp

1a. ok-voon ororok sprok .	7a. lalok farok ororok lalok sprok izok enemok .
1b. at-voon bichat dat .	7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok .	8a. lalok brok anok plok nok .
2b. at-drubel at-voon pippat rrat dat .	8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok .	9a. wiwok nok izok kantok ok-yurp .
3b. totat dat arrat vat hilat .	9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok .	10a. lalok mok nok yorok ghirok clok .
4b. at-voon krat pippat sat lat .	10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok .	11a. lalok nok crrrok hihok yorok zanzanok .
5b. totat jjat quat cat .	11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok .	12a. lalok rarok nok izok hihok mok .
6b. wat dat krat quat cat .	12b. wat nnat forat arrat vat gat .

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: farok crrrok hihok yorok clok kantok ok-yurp

1a. ok-voon ororok sprok .	7a. lalok farok ororok lalok sprok izok enemok .
1b. at-voon bichat dat .	7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok .	8a. lalok brok anok plok nok .
2b. at-drubel at-voon pippat rrat dat .	8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok .	9a. wiwok nok izok kantok ok-yurp .
3b. totat dat arrat vat hilat .	9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok .	10a. lalok mok nok yorok ghirok clok .
4b. at-voon krat pippat sat lat .	10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok .	11a. lalok nok crrrok hihok yorok zanzanok .
5b. totat jjat quat cat .	11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok .	12a. lalok rarok nok izok hihok mok .
6b. wat dat krat quat cat .	12b. wat nnat forat arrat vat gat .

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: farok crrok hihok yorok clok kantok ok-yurp

1a. ok-voon ororok sprok .	7a. lalok farok ororok lalok sprok izok enemok . /
1b. at-voon bichat dat .	7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok .	8a. lalok brok anok plok nok .
2b. at-drubel at-voon pippat rrat dat .	8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok .	9a. wiwok nok izok kantok ok-yurp .
3b. totat dat arrat vat hilat .	9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok .	10a. lalok mok nok yorok ghirok clok .
4b. at-voon krat pippat sat lat .	10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok .	11a. lalok nok crrok hihok yorok zanzanok .
5b. totat jjat quat cat .	11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok .	12a. lalok rarok nok izok hihok mok .
6b. wat dat krat quat cat .	12b. wat nnat forat arrat vat gat .

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: farok crrrok hihok yorok clok kantok ok-yurp

1a. ok-voon ororok sprok .	7a. lalok farok ororok lalok sprok izok enemok . /
1b. at-voon bichat dat .	7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok .	8a. lalok brok anok plok nok .
2b. at-drubel at-voon pippat rrat dat .	8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok .	9a. wiwok nok izok kantok ok-yurp .
3b. totat dat arrat vat hilat .	9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok .	10a. lalok mok nok yorok ghirok clok .
4b. at-voon krat pippat sat lat .	10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . /	11a. lalok nok crrrok hihok yorok zanzanok . ???
5b. totat jjat quat cat .	11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok .	12a. lalok rarok nok izok hihok mok .
6b. wat dat krat quat cat .	12b. wat nnat forat arrat vat gat .

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: farok crrrok hihok yorok clok kantok ok-yurp

1a. ok-voon ororok sprok .	7a. lalok farok ororok lalok sprok izok enemok . /
1b. at-voon bichat dat .	7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok .	8a. lalok brok anok plok nok .
2b. at-drubel at-voon pippat rrat dat .	8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok .	9a. wiwok nok izok kantok ok-yurp .
3b. totat dat arrat vat hilat .	9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok .	10a. lalok mok nok yorok ghirok clok .
4b. at-voon krat pippat sat lat .	10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . /	11a. lalok nok crrrok hihok yorok zanzanok .
5b. totat jjat quat cat .	11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok .	12a. lalok rarok nok izok hihok mok .
6b. wat dat krat quat cat .	12b. wat nnat forat arrat vat gat .

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: farok crrrok **hihok** yorok clok kantok ok-yurp

1a. ok-voon ororok sprok .	7a. lalok farok ororok lalok sprok izok enemok . /
1b. at-voon bichat dat .	7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok .	8a. lalok brok anok plok nok .
2b. at-drubel at-voon pippat rrat dat .	8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok .	9a. wiwok nok izok kantok ok-yurp .
3b. totat dat arrat vat hilat .	9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok .	10a. lalok mok nok yorok ghirok clok .
4b. at-voon krat pippat sat lat .	10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok .	11a. lalok nok crrrok hihok yorok zanzanok .
5b. totat jjat quat cat .	11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok .	12a. lalok rarok nok izok hihok mok .
6b. wat dat krat quat cat .	12b. wat nnat forat arrat vat gat .

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: farok crrrok hihok yorok **clok** kantok ok-yurp

1a. ok-voon ororok sprok .	7a. lalok farok ororok lalok sprok izok enemok . /
1b. at-voon bichat dat .	7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok .	8a. lalok brok anok plok nok .
2b. at-drubel at-voon pippat rrat dat .	8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok .	9a. wiwok nok izok kantok ok-yurp .
3b. totat dat arrat vat hilat .	9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok .	10a. lalok mok nok yorok ghirok clok .
4b. at-voon krat pippat sat lat .	10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok .	11a. lalok nok crrrok hihok yorok zanzanok .
5b. totat jjat quat cat .	11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok .	12a. lalok rarok nok izok hihok mok .
6b. wat dat krat quat cat .	12b. wat nnat forat arrat vat gat .

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: farok crrrok hihok yorok **clok** kantok ok-yurp

1a. ok-voon ororok sprok .	7a. lalok farok ororok lalok sprok izok enemok . /
1b. at-voon bichat dat .	7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok .	8a. lalok brok anok plok nok .
2b. at-drubel at-voon pippat rrat dat .	8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok .	9a. wiwok nok izok kantok ok-yurp .
3b. totat dat arrat vat hilat .	9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok .	10a. lalok mok nok yorok ghirok clok . / ???
4b. at-voon krat pippat sat lat .	10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok .	11a. lalok nok crrrok hihok yorok zanzanok . /
5b. totat jjat quat cat .	11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok .	12a. lalok rarok nok izok hihok mok . /
6b. wat dat krat quat cat .	12b. wat nnat forat arrat vat gat .

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: farok crrrok hihok yorok clok kantok ok-yurp

1a. ok-voon ororok sprok .	7a. lalok farok ororok lalok sprok izok enemok . /
1b. at-voon bichat dat .	7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok .	8a. lalok brok anok plok nok . /
2b. at-drubel at-voon pippat rrat dat .	8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok .	9a. wiwok nok izok kantok ok-yurp .
3b. totat dat arrat vat hilat .	9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok .	10a. lalok mok nok yorok ghirok clok . X / /
4b. at-voon krat pippat sat lat .	10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok .	11a. lalok nok crrrok hihok yorok zanzanok . / / /
5b. totat jjat quat cat .	11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok .	12a. lalok rarok nok izok hihok mok . / / /
6b. wat dat krat quat cat .	12b. wat nnat forat arrat vat gat .

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: farok crrrok hihok yorok **clok** kantok ok-yurp

1a. ok-voon ororok sprok .	7a. lalok farok ororok lalok sprok izok enemok . /
1b. at-voon bichat dat .	7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok .	8a. lalok brok anok plok nok . /
2b. at-drubel at-voon pippat rrat dat .	8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok .	9a. wiwok nok izok kantok ok-yurp .
3b. totat dat arrat vat hilat .	9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok .	10a. lalok mok nok yorok ghirok clok . X / •••••••• process of elimination
4b. at-voon krat pippat sat lat .	10b. wat nnat gat mat •••••••• bat hilat .
5a. wiwok farok izok stok .	11a. lalok nok crrrok hihok yorok zanzanok . / ••••••••
5b. totat jjat quat cat .	11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok .	12a. lalok rarok nok izok hihok mok . •••••••• / ••••••••
6b. wat dat krat quat cat .	12b. wat nnat forat arrat vat gat .

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: farok crrrok hihok yorok clok kantok ok-yurp

1a. ok-voon ororok sprok .	7a. lalok farok ororok lalok sprok izok enemok . /
1b. at-voon bichat dat .	7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok .	8a. lalok brok anok plok nok . /
2b. at-drubel at-voon pippat rrat dat .	8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok .	9a. wiwok nok izok kantok ok-yurp .
3b. totat dat arrat vat hilat .	9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok .	10a. lalok mok nok yorok ghirok clok . X /
4b. at-voon krat pippat sat lat .	10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok .	11a. lalok nok crrrok hihok yorok zanzanok . /
5b. totat jjat quat cat .	11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok .	12a. lalok rarok nok izok hihok mok . /
6b. wat dat krat quat cat .	12b. wat nnat forat arrat vat gat .

Centauri/Arcturan [Knight, 1997]

Your assignment, put these words in order: { jjat, arrat, mat, bat, oloat, at-yurp }

1a. ok-voon ororok sprok .	7a. lalok farok ororok lalok sprok izok enemok . /
1b. at-voon bichat dat .	7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok .	8a. lalok brok anok plok nok . /
2b. at-drubel at-voon pippat rrat dat .	8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok .	9a. wiwok nok izok kantok ok-yurp .
3b. totat dat arrat vat hilat .	9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok .	10a. lalok mok nok yorok ghirok clok . X / X /
4b. at-voon krat pippat sat lat .	10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok .	11a. lalok nok crrok hihok yorok zanzanok . / cr / z / zero / fertility
5b. totat jjat quat cat .	11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok .	12a. lalok rarok nok izok hihok mok . / / /
6b. wat dat krat quat cat .	12b. wat nnat forat arrat vat gat .

Centauri/Arcturan [Knight, 1997]

Your assignment, put these words in order:
{ jjat, arrat, mat, bat, oloat, at-yurp }

- There are $6!$ different orders possible, so 720 different translations.
- Best order (according to placement in TL side of the corpus is as given above):
 - Not just unigrams, but n -grams also ...

It's Really Spanish—English!

Clients do not sell pharmaceuticals in Europe => Clientes no venden medicinas en Europa

1a. Garcia and associates .
1b. Garcia y asociados .

7a. the clients and the associates are enemies .
7b. los clientes y los asociados son enemigos .

2a. Carlos Garcia has three associates .
2b. Carlos Garcia tiene tres asociados .

8a. the company has three groups .
8b. la empresa tiene tres grupos .

3a. his associates are not strong .
3b. sus asociados no son fuertes .

9a. its groups are in Europe .
9b. sus grupos estan en Europa .

4a. Garcia has a company also .
4b. Garcia tambien tiene una empresa .

10a. the modern groups sell strong pharmaceuticals .
10b. los grupos modernos venden medicinas fuertes .

5a. its clients are angry .
5b. sus clientes estan enfadados .

11a. the groups do not sell zenzanine .
11b. los grupos no venden zanzanina .

6a. the associates are also angry .
6b. los asociados tambien estan enfadados .

12a. the small groups are not modern .
12b. los grupos pequenos no son modernos .

Some more to try ...

- iat lat pippat eneat hilat oloat at-yurp.
- totat nnat forat arrat mat bat.
- wat dat quat cat uskrat at-drubel.

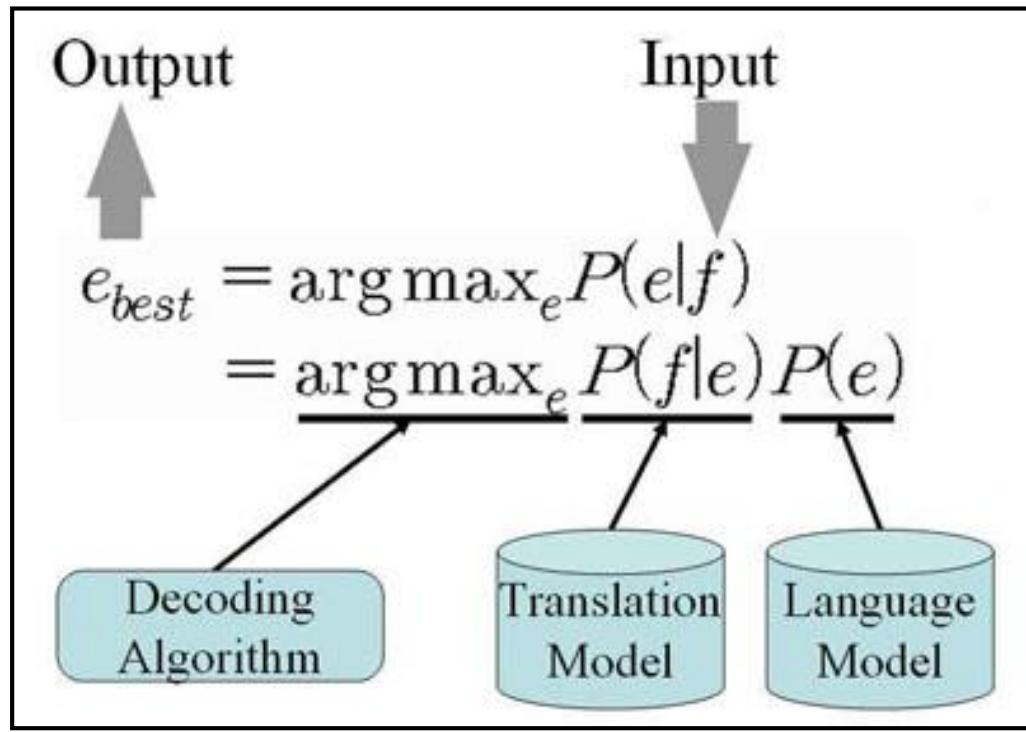
Some more to try ...

- iat lat pippat eneat hilat oloat at-yurp
- totat nnat forat arrat mat bat.
- wat dat quat cat uskrat at-drubel.

... if you have trouble sleeping at night!



How does SMT Work?



- No(t much) maths today ... 😊
- Instead:
 - The story of SMT in pictures ...
 - It's (mostly) all about the **Data** ...

How does SMT Work?

Statistical MT learns from data

Two kinds of data:

- Source documents and their human translations
 - Target language collections
-
- The more data the better!
 - Also: the **right kind** of data!

GERMAN	ENGLISH	FRENCH
<p>Einleitung</p> <p><i>I. Von dem Unterschiede der reinen und empirischen Erkenntnis</i></p> <p>Daß alle unsere Erkenntnis mit der Erfahrung anfange, daran ist gar kein Zweifel; denn wodurch sollte das Erkenntnisvermögen sonst zur Ausübung erweckt werden, geschähe es nicht durch Gegenstände, die unsere Sinne rühren und teils von selbst Vorstellungen bewirken, teils unsere Verständigkeit in Bewegung bringen, diese zu vergleichen, sie zu verknüpfen oder zu trennen, und so den rohen Stoff sinnlicher Eindrücke zu einer Erkenntnis der Gegenstände zu verarbeiten, die Erfahrung heißt? Der Zeit nach geht also keine Erkenntnis in uns vor der Erfahrung vorher, und mit dieser fängt alle an.</p>	<p>Introduction</p> <p><i>I. Of the difference between Pure and Empirical Knowledge</i></p> <p>That all our knowledge begins with experience there can be no doubt. For how is it possible that the faculty of cognition should be awakened into exercise otherwise than by means of objects which affect our senses, and partly of themselves produce representations, partly rouse our powers of understanding into activity, to compare to connect, or to separate these, and so to convert the raw material of our sensuous impressions into a knowledge of objects, which is called experience? In respect of time, therefore, no knowledge of ours is antecedent to experience, but begins with it.</p>	<p>Introduction</p> <p><i>I. De la différence de la connaissance pure et de la connaissance empirique.</i></p> <p>Que toute notre connaissance commence avec l'expérience, cela ne soulève aucun doute. En effet, par quoi notre pouvoir de connaître pourrait-il être éveillé et mis en action, si ce n'est par des objets qui frappent nos sens et qui, d'une part, produisent par eux-mêmes des représentations et, d'autre part, mettent en mouvement notre faculté intellectuelle, afin qu'elle compare, lie ou sépare ces représentations, et travaille ainsi la matière brute des impressions sensibles pour en tirer une connaissance des objets, celle qu'on nomme l'expérience? Ainsi, chronologiquement, aucune connaissance ne précède en nous l'expérience et c'est avec elle que toutes commencent.</p>

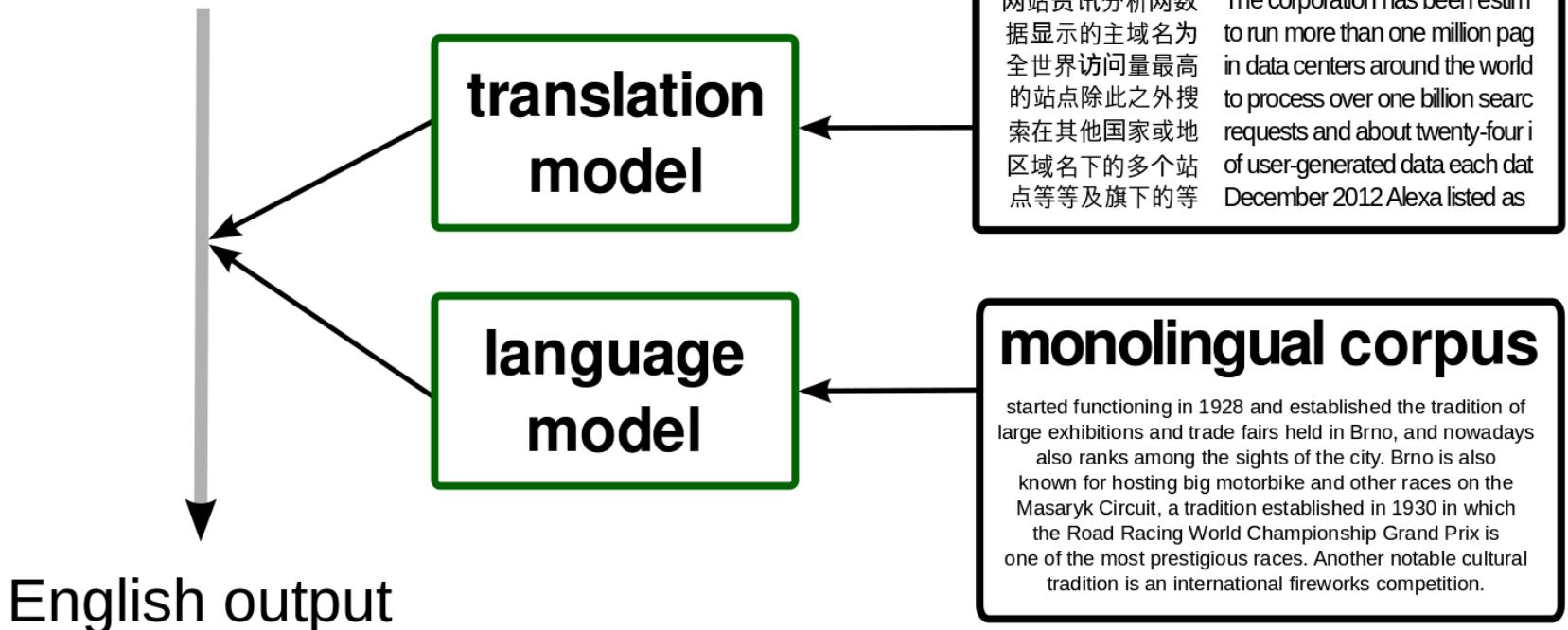
What can/do we learn from Data?

- Which sentences translate as which: **sentence alignment**
- Which words translate as which: **word alignment + translation probabilities**
=> **translation model**
- What do good target sentences look like:
language model

GERMAN	ENGLISH	FRENCH
<p>Einleitung</p> <p><i>I. Von dem Unterschiede der reinen und empirischen Erkenntnis</i></p> <p>Daß alle unsere Erkenntnis mit der Erfahrung anfange, daran ist gar kein Zweifel; denn wodurch sollte das Erkenntnisvermögen sonst zur Ausübung erweckt werden, geschähe es nicht durch Gegenstände, die unsere Sinne rühren und teils von selbst Vorstellungen bewirken, teils unsere Verständigkeit in Bewegung bringen, diese zu vergleichen, sie zu verknüpfen oder zu trennen, und so den rohen Stoff sinnlicher Eindrücke zu einer Erkenntnis der Gegenstände zu verarbeiten, die Erfahrung heißt? Der Zeit nach geht also keine Erkenntnis in uns vor der Erfahrung vorher, und mit dieser fängt alle an.</p>	<p>Introduction</p> <p><i>I. Of the difference between Pure and Empirical Knowledge</i></p> <p>That all our knowledge begins with experience there can be no doubt. For how is it possible that the faculty of cognition should be awakened into exercise otherwise than by means of objects which affect our senses, and partly of themselves produce representations, partly rouse our powers of understanding into activity, to compare to connect, or to separate these, and so to convert the raw material of our sensuous impressions into a knowledge of objects, which is called experience? In respect of time, therefore, no knowledge of ours is antecedent to experience, but begins with it.</p>	<p>Introduction</p> <p><i>I. De la différence de la connaissance pure et de la connaissance empirique.</i></p> <p>Que toute notre connaissance commence avec l'expérience, cela ne soulève aucun doute. En effet, par quoi notre pouvoir de connaître pourrait-il être éveillé et mis en action, si ce n'est par des objets qui frappent nos sens et qui, d'une part, produisent par eux-mêmes des représentations et, d'autre part, mettent en mouvement notre faculté intellectuelle, afin qu'elle compare, lie ou sépare ces représentations, et travaille ainsi la matière brute des impressions sensibles pour en tirer une connaissance des objets, celle qu'on nomme l'expérience? Ainsi, chronologiquement, aucune connaissance ne précède en nous l'expérience et c'est avec elle que toutes commencent.</p>

How does SMT Work?

似乎格式有問題



Sentence Alignment

TRADOS WinAlign - [C:\...\VOW_E_H_Ex_wo_mod.dll : C:\...\VOW_D_H_Ex_wo_mod.dll] [Word Documents]

File Edit View Settings Segment Window Help

C:\...\VOW_E_H_Ex_wo_mod.dll
CWMatch

This brings you to the calendar where you can begin to enter data.

After indicating your first cycle date in the dialog box, a calendar appears.

It indicates the month of your first cycle date.

The day you indicated as the first cycle day will be marked with a ^g1 in the lower right hand corner and a ^g2.

If this is not the day you want to indicate as the first cycle day, go to File, Close, and don't save it.

Then double-click on the CycleMatch icon, enter the correct date, and click on OK.

You will also see two numbers in each calendar day.

The number in the upper right Ready.

C:\...\VOW_D_H_Ex_wo_mod.dll
CWMatch

Dadurch gelangen Sie zum Kalender, in dem alle weiteren Eingaben erfolgen.

Nachdem Sie Ihren ersten Zyklus in das Dialogfenster eingegeben haben erscheint eine Kalenderansicht, die den Monat Ihres ersten Zyklus wiedergibt.

Der Tag, den Sie als ersten Zyklustag eingegeben haben, wird in der rechten unteren Ecke des Kalenderfensters durch ^g1 und ^g2 dargestellt.

Falls dies nicht der erste Tag Ihres ersten Zyklus ist, wählen Sie bitte "Datei", "Schließen", und "Nicht Speichern".

Führen Sie dann bitte erneut einen Doppelklick auf das CycleMatch Icon aus und korrigieren das Datum.

Um die Eingabe zu bestätigen klicken Sie bitte auf OK.

Es werden jeweils zwei Zahlen pro Kalendertag dargestellt.

In der nächsten Folie oben unten

NUM

The screenshot shows a dual-pane interface for sentence alignment. The left pane displays the English source text, and the right pane displays the German target text. Lines connect specific text elements from the English side to their corresponding translations on the German side, illustrating the alignment process. The English text describes the CycleMatch feature, while the German text provides step-by-step instructions for using it.

Let's try it for a
language pair that
someone in the class
might know ...

Word Alignment

清 燉 雞
雞 飯
雞 麵
廣 東 雲 吞
蕃 茄 香
雲 吞
酸 辣
香
雲 香
豆 腐 菜
雞 玉 米
蟹 肉 玉 米
海 鮮

CLASSIC SOUPS

			Sm.	Lg.
	57.	House Chicken Soup (Chicken, Celery, Potato, Onion, Carrot)	1.50	2.75
	58.	Chicken Rice Soup	1.85	3.25
	59.	Chicken Noodle Soup	1.85	3.25
	60.	Cantonese Wonton Soup.....	1.50	2.75
	61.	Tomato Clear Egg Drop Soup	1.65	2.95
	62.	Regular Wonton Soup	1.10	2.10
	63.	Hot & Sour Soup	1.10	2.10
	64.	Egg Drop Soup.....	1.10	2.10
	65.	Egg Drop Wonton Mix.....	1.10	2.10
	66.	Tofu Vegetable Soup	NA	3.50
	67.	Chicken Corn Cream Soup	NA	3.50
	68.	Crab Meat Corn Cream Soup.....	NA	3.50
	69.	Seafood Soup.....	NA	3.50

Word Alignment

CLASSIC SOUPS				Sm.	Lg.
清 燉 雞 湯	57.	House Chicken Soup (Chicken, Celery, Potato, Onion, Carrot)		1.50	2.75
雞 飯 湯	58.	Chicken Rice Soup		1.85	3.25
雞 麵 湯	59.	Chicken Noodle Soup		1.85	3.25
廣 東 雲 吞	60.	Cantonese Wonton Soup.....		1.50	2.75
蕃 茄 香 湯	61.	Tomato Clear Egg Drop Soup		1.65	2.95
雲 吞 湯	62.	Regular Wonton Soup		1.10	2.10
酸 辣 湯	63.	Hot & Sour Soup		1.10	2.10
香 花 湯	64.	Egg Drop Soup.....		1.10	2.10
雲 香 湯	65.	Egg Drop Wonton Mix.....		1.10	2.10
豆 腐 菜 湯	66.	Tofu Vegetable Soup	NA	3.50	
雞 玉 米 湯	67.	Chicken Corn Cream Soup	NA	3.50	
蟹 肉 玉 米 湯	68.	Crab Meat Corn Cream Soup.....	NA	3.50	
海 鮮 湯	69.	Seafood Soup.....	NA	3.50	

Let's try it for a
language pair that
most of you
in the class know ...

Statistical MT

I love the boy.
J'aime le garçon.

I love the dog.
J'aime le chien.

They love the dog.
Ils aiment le chien.

They talk to the girl.
Ils parlent à la fille.

They talk to the dog.
Ils parlent au chien.

I talk to the mother.
Je parle à la mère.

Aligned Data

Statistical MT

I love the boy.
J'aime le garçon.

I love the dog.
J'aime le chien.

They love the dog.
Ils aiment le chien.

They talk to the girl.
Ils parlent à la fille.

They talk to the dog.
Ils parlent au chien.

I talk to the mother.
Je parle à la mère.



I	J'	II	mother	mère	1
	Je	I	dog	chien.	III
love	aime	II	they	ils	III
	aiment	I	talk	parlent	II
the	le	III		parle	1
	la	II	to	à	II
boy	garçon	I		au/_the	1
girl	fille	I			

Collated Statistics

Aligned Data

Statistical MT

I love the boy.
J'aime le garçon.

I love the dog.
J'aime le chien.

They love the dog.
Ils aiment le chien.

They talk to the girl.
Ils parlent à la fille.

They talk to the dog.
Ils parlent au chien.

I talk to the mother.
Je parle à la mère.



Input

I talk to the girl.

I	J'	II	mother	mère	I
	Je	I	dog	chien.	III
love	aime	II	they	ils	III
	aiment	I	talk	parlent	II
the	le	III		parle	I
	la	II	to	à	II
boy	garçon	I		au/_the	I
girl	fille	I			

Collated Statistics

Aligned Data

Statistical MT

I love the boy.
J'aime le garçon.

I love the dog.
J'aime le chien.

They love the dog.
Ils aiment le chien.

They talk to the girl.
Ils parlent à la fille.

They talk to the dog.
Ils parlent au chien.

I talk to the mother.
Je parle à la mère.

Aligned Data



Input

I talk to the girl.

I	J'	II	mother	mère	I
	Je	I	dog	chien	III
love	aime	II	they	ils	III
	aiment	I	talk	parlent	II
the	le	III		parle	I
	la	II	to	à	II
boy	garçon	I		au/_the	I
girl	fille	I			

Collated Statistics

+

Language Model

J'parlent à la fille.

Output

Statistical MT

I love the boy.
J'aime le garçon.
I love the dog.
J'aime le chien.
They love the dog.
Ils aiment le chien.
They talk to the girl.
Ils parlent à la fille.
They talk to the dog.
Ils parlent au chien.
I talk to the mother.
Je parle à la mère.



Aligned Data

I	talk	to	the	girl
J'	parlent		au	le fille
2/3	2/3		2/3	3/5 1/1
Je	parle	à	la	fille
1/3	1/3	1/3	2/5	1/1

How to choose?

Statistical MT



The Language Model:

- What is good target language?
- Which words can follow which words and which can't? The “grammar”!
- Learnt from the data ...
 - Je parle is good ...
 - J' parlent is bad ...
 - la fille is good ...
 - le fille is bad ...
- Je parle à la fille >> J' parlent à la fille

Phrase-Based SMT

- So far: translating single words
- Loses context: such as agreement (**le fille* ...) etc.
- To some extent “**repaired**” by language model
- A better model:
 - Not just translations of single words
 - But also **phrase** translations:

*– the girl : la fille
– to the girl : à la fille
– I talk : Je parle*

Phrase-Based SMT

I love the boy.
J'aime le garçon.

I love the dog.
J'aime le chien.

They love the dog.
Ils aiment le chien.

They talk to the girl.
Ils parlent à la fille.

They talk to the dog.
Ils parlent au chien.

I talk to the mother.
Je parle à la mère.



Aligned Data

Input

I talk to the girl.

I love	J'aime	11
They love	Ils aiment	1
They talk	Ils parlent	11
I talk	Je parle	1
To the dog	au chien	1
the boy	le garçon	1
the dog	le chien	11
to the girl	à la fille	1
to the boy	au garçon	1
to the mother	à la mère	1

Phrase-Based SMT

I love the boy.
J'aime le garçon.

I love the dog.
J'aime le chien.

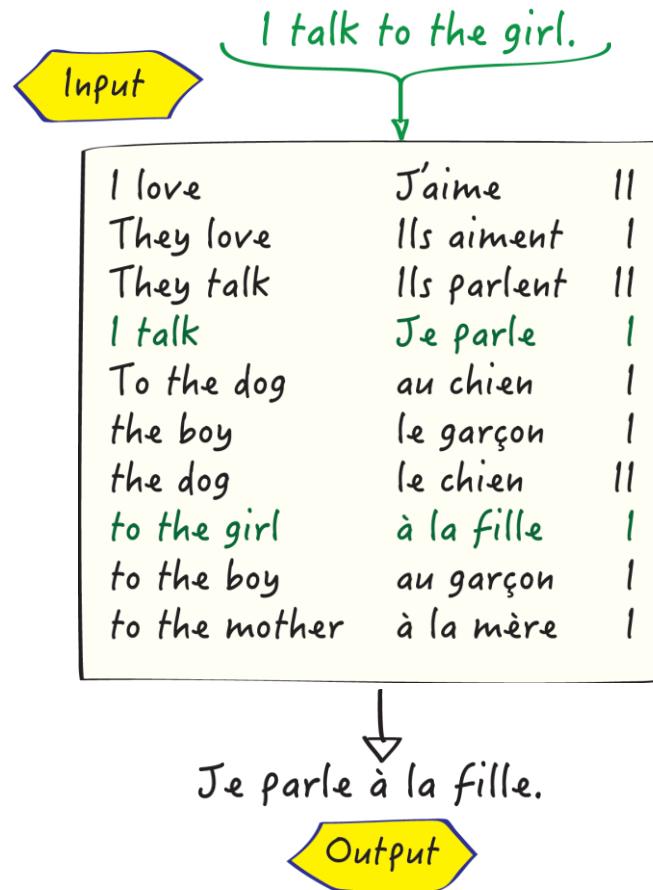
They love the dog.
Ils aiment le chien.

They talk to the girl.
Ils parlent à la fille.

They talk to the dog.
Ils parlent au chien.

I talk to the mother.
Je parle à la mère.

Aligned Data



Phrase-Based SMT

- *Much* better than word-based SMT!
- (Was) Standard technology: Google, Microsoft, Baidu, global Localisation & Translation industry
- Moses open-source PB-SMT toolkit
- Most widely used SMT platform
- Research funded by EC
- (Was) used by EC DGT's MT@EC



What did we learn?

- what parallel corpora look like (more on this soon);
- viewing parallel corpora through the ‘eyes’ of a computer;
- how relevant parallel corpora are for MT;
- how to build bilingual dictionaries from parallel corpora;
- how cognate information may be useful in MT;
- how to do word alignment;
- about the ‘chicken-and-egg’ nature of dictionaries (which enable word alignments) and word alignments (which enable dictionary writing) ...

What else do we need to know?

- about word alignment and dictionary writing on a larger scale;
- about phrasal alignment, the norm in real translation data;
- about unalignable words;
- the importance of knowing the target language (vs. source) in making fluent translations;
- the importance of short sentence pairs (where alignment possibilities are restricted) in helping disambiguate/align longer sentence pairs;
- about locality in word order shifts;
- how to guess the meanings/translations of unknown words;
- about how much uncertainty the machine faces in working with limited data;
- ...

Can such methods be scaled to ‘real’ MT?

- Availability of monolingual and bilingual corpora?
- Possibility of sentence-aligning bilingual corpora?
- Can we write an algorithm to extract the translation dictionary?
- Can we write an algorithm to extract the monolingual word pair counts?
- Can we write an algorithm to generate translations using our translation dictionary and word pair counts?
- **WILL THE TRANSLATIONS PRODUCED BE ANY GOOD?**

Parallel Corpora

- Hugely important ... but not available in a wide range of language pairs:
 - Chinese—English: Hong Kong data
 - French—English: Canadian Hansards
 - Older EU pairs: Europarl [Koehn 04]
 - Newer EU pairs: JRC-Acquis Communautaire
 - Arabic—English: LDC Data
 - NIST, IWSLT, TC-STAR Evaluations

Good Quality Language & Translation Models

- Any statistical approach to MT requires the availability of aligned bilingual corpora which are:
 - large;
 - good-quality;
 - representative.

Corpus 1

Mary and John have two children.
The children that Mary and John have are aged 3 and 4.
John has blue eyes.

Question 1: what's $P(\text{have})$ vs. $P(\text{has})$ in a corpus?

Question 2: what's $P(\text{have} \mid \text{John})$ vs. $P(\text{has} \mid \text{John})$ in a corpus?

Question 3: what's $P(\text{have})$ vs. $P(\text{has})$ in *this* corpus? What's their *relative* probability?

Question 4: what's $P(\text{have} \mid \text{John})$ vs. $P(\text{has} \mid \text{John})$ in *this* corpus?

Corpus 2

Am I right, or am I wrong?
Peter and I are seldom wrong.
I am sometimes right.
Sam and I are often mistaken.

Question 5: What two generalisations would a probabilistic language model (based on *bigrams*, say) infer from this data, which are not true of English as a whole? Are there any other generalisations that could be inferred?

Question 6: Try to think of some trigrams (and 4-grams, if you can) that cannot be ‘discovered’ by a bigram model? What you’re looking for here is a phrase where the third (or subsequent) word depends on the first word, which in a bigram model is ‘too far away’ ...

Some Observations

- Note that all the sentences in these corpora are well-formed.
- If, on the other hand, the corpus contains ill-formed input, then that too will skew our probability models ...

... and our translations will be affected!

Corpus 1 Revisited

- Using Google, I got:
 - # ‘have’ = 380,000,000
 - # ‘has’ = 244,000,000
 - # ‘John has’ = 227,000
 - # ‘John have’ = 25,700
- Revisit the Questions and calculate the *actual* probabilities! How accurate/inaccurate were the original models that we derived?

Corpus 2 Revisited

- Using Google, I got:
 - # ‘am I’ = 3,690,000
 - # ‘I am’ = 8,060,000
 - # ‘I are’ = 1,230,000
- Revisit the Questions and calculate the *actual* probabilities! How accurate/inaccurate were the original models that we derived?

Bilingual Corpora

All this applies to
bitexts too!

Q: of what English word
are these possible
French translations
(from the *Canadian
Hansards*, note)?

Q: what's ???

French	Probability
???	.808
entendre	.079
entendu	.026
entends	.024
entendons	.013

Caveat interpres!

- Beware of sparse data!
- Beware of unrepresentative corpora!
- Beware of poor quality language!

If the corpora are small, or of poor quality, or are unrepresentative, then our statistical language models will be poor, so any results we achieve will be poor.

Noisy Channel Framework

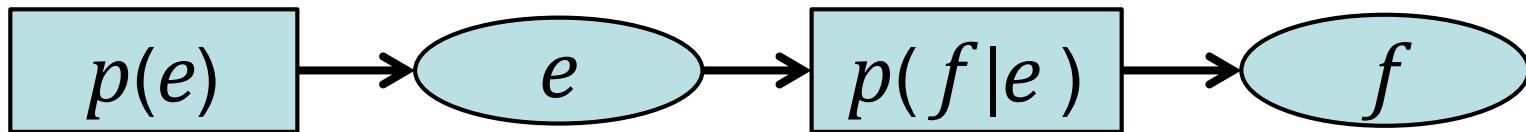


English



French

Noisy Channel Framework

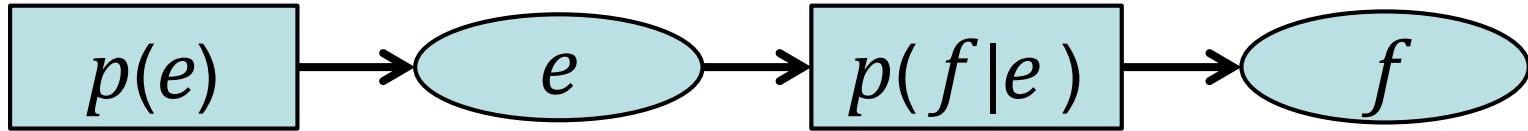


English



French

Noisy Channel Framework



English

$$p(f) = p(e)p(f|e)$$



French

Noisy Channel Framework

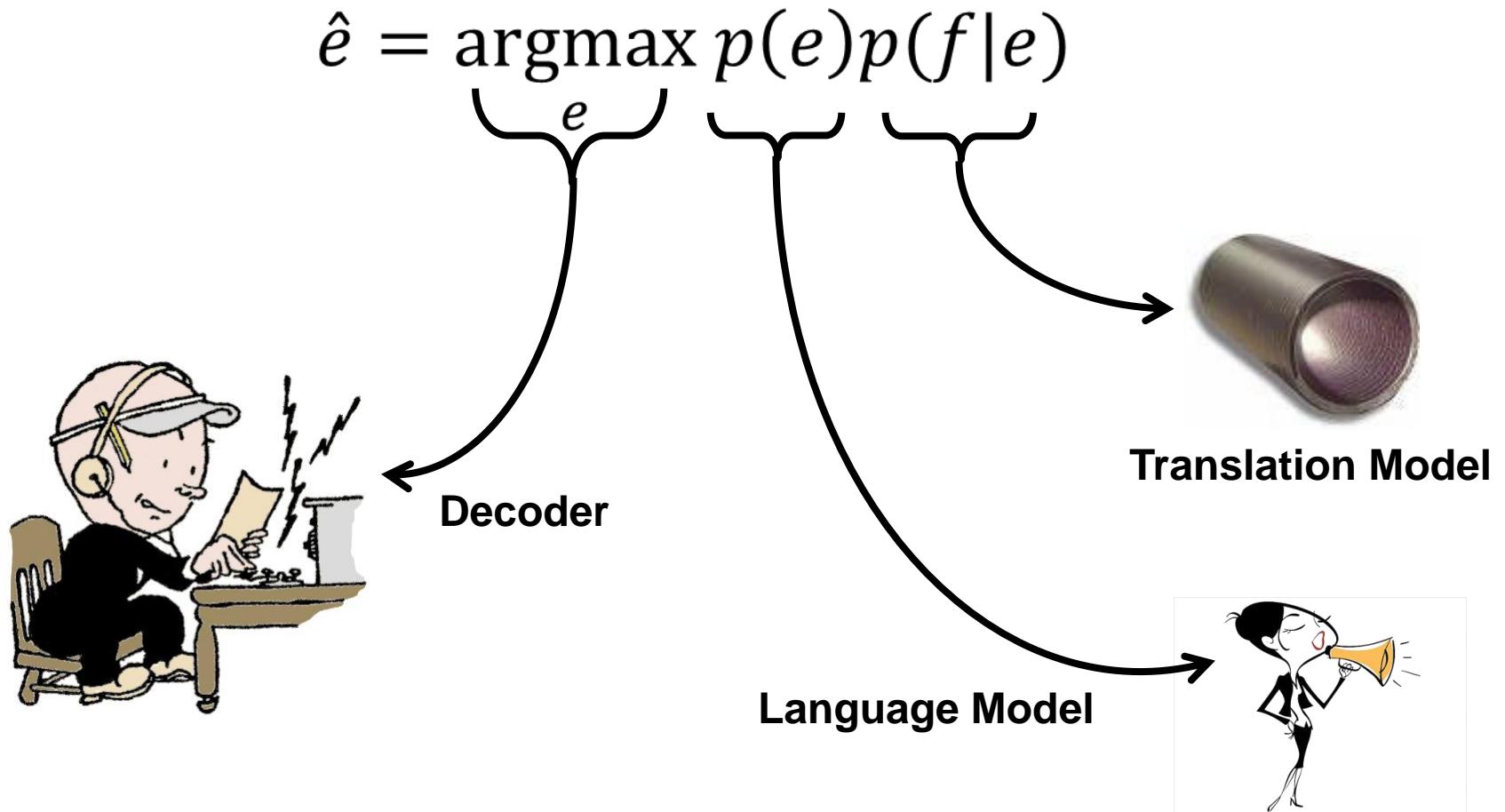
Applying Bayes' Rule, we have:

$$p(e|f) = \frac{p(e)p(f|e)}{p(f)}$$

Thus:

$$\hat{e} = \operatorname{argmax}_e p(e|f) = \operatorname{argmax}_e p(e)p(f|e)$$

SMT Components



Noisy Channel Framework

- The *translation model* models how likely it is that f is a translation of e – **adequacy**.
- The *language model* models how likely it is that e is an acceptable sentence – **fluency**.
- The *decoder* searches for the most likely e .

Fluency versus Adequacy

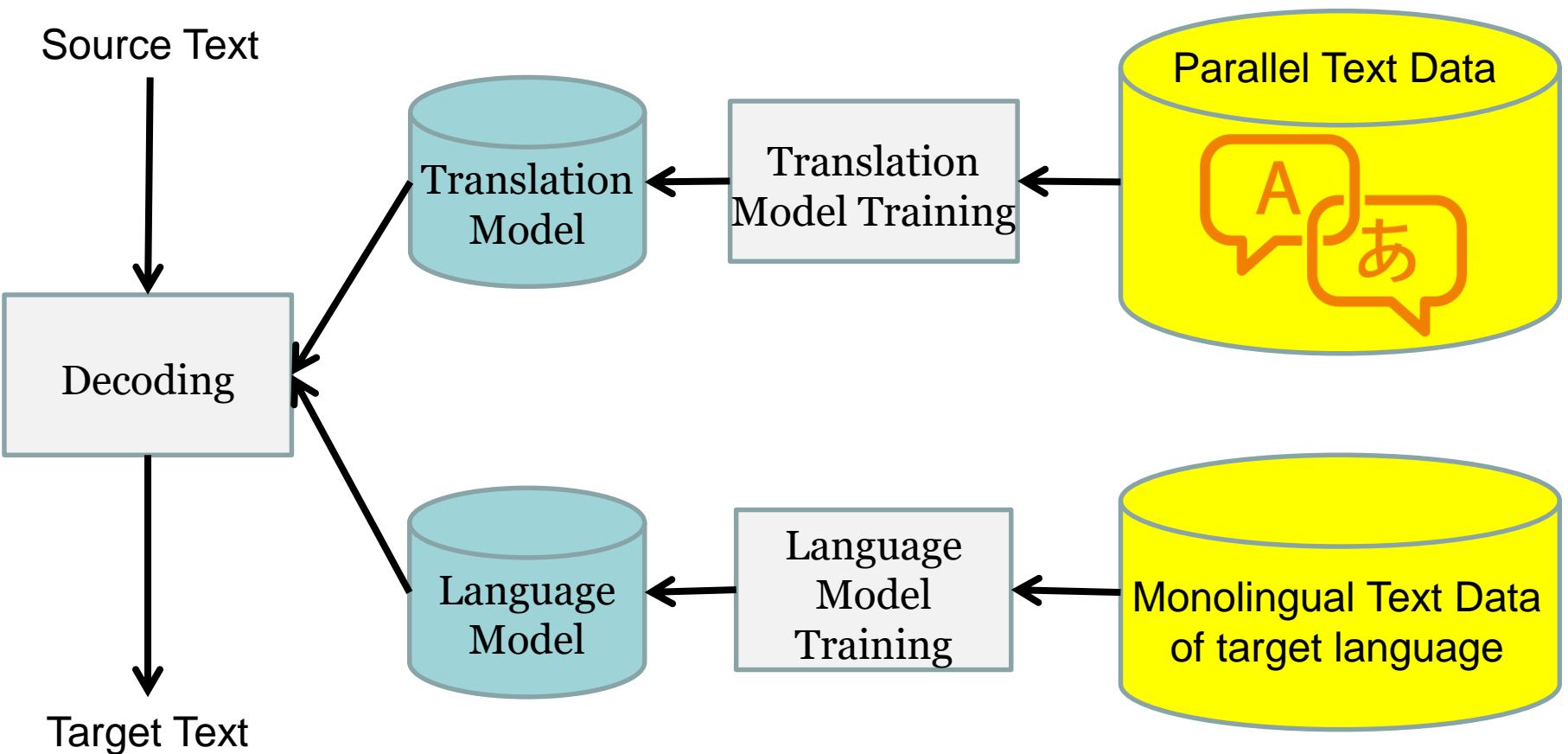
(towards MT Evaluation)

Source Sentence:

Le chat entre dans la chambre.

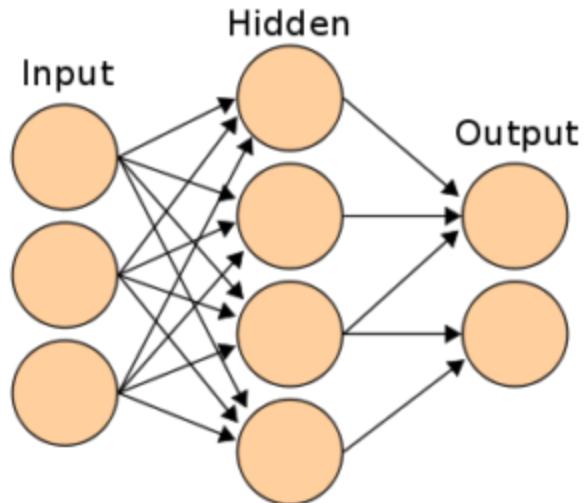
- Adequate fluent translation:
The cat enters the bedroom.
- Adequate disfluent translation:
The cat enters in the bedroom.
- Fluent inadequate translation:
My Granny plays the piano.
- Disfluent inadequate translation:
piano Granny the plays My

SMT Flow

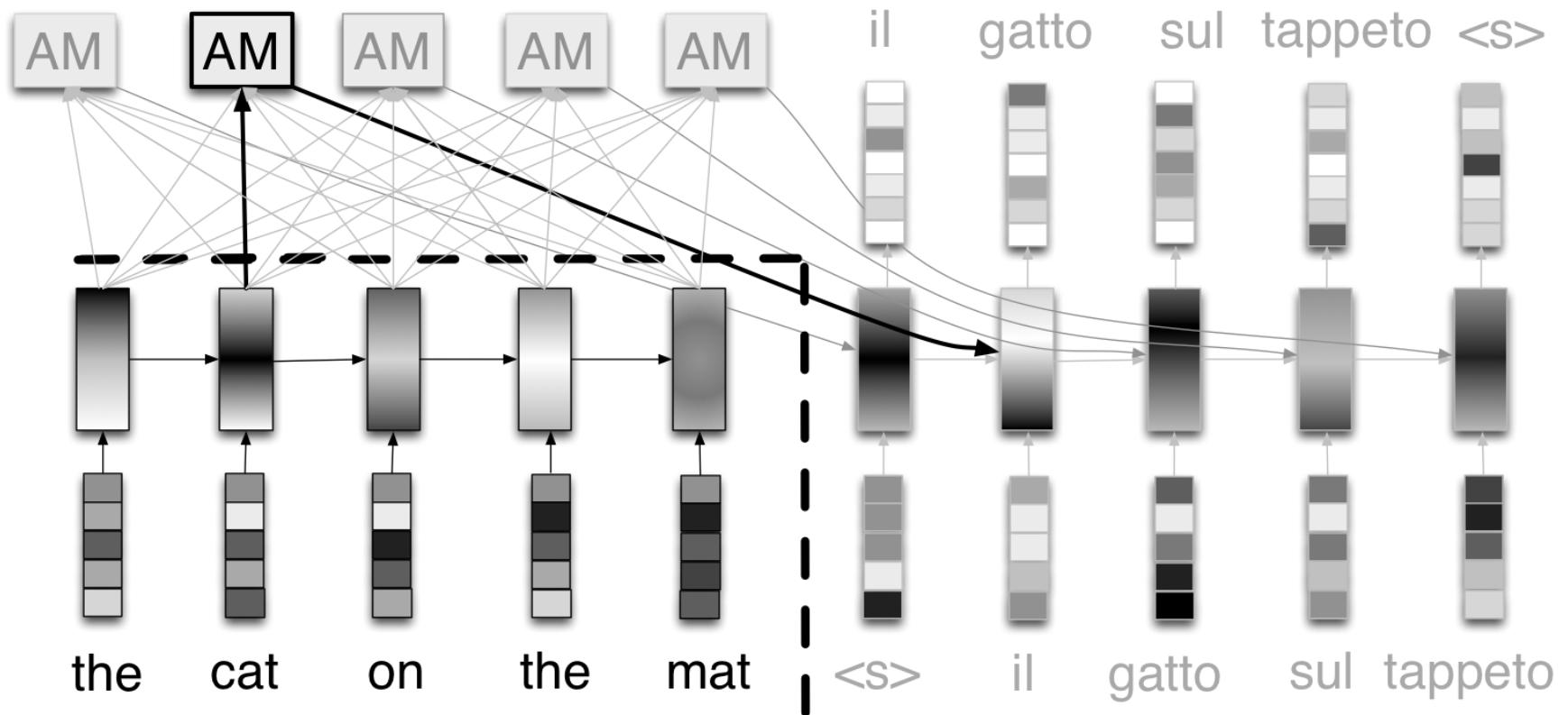


Neural MT

- Paradigm Shift:
Machine Learning →
Deep Learning
- Hardware: CPU → GPU
- Open-Source Tools:
 - *OpenNMT*
 - *Tensorflow*
- Slower Training &
Decoding



SOTA in NMT: Encoder-Decoder with Attention



Neural MT

- Sequence-to-sequence translation, not word-to-word (or phrase-to-phrase)
- Very promising results compared to PB-SMT:
 - Good generalisation capability
 - Good predictive performance
- But some (increasingly disappearing) teething problems: issues with sentence length, vocab size

Neural MT

- Sequence-to-sequence translation, not word-to-word (or phrase-to-phrase)
- Very promising results compared to PB-SMT:
 - Good generalisation capability
 - Good predictive performance
- But some (increasingly disappearing) teething problems: issues with sentence length, vocab size

Is NMT the new SOTA?

OK, so that's set the scene

I hope that's enough to
get you
started/interested in
SMT (and then NMT) ...