

Speech and Natural Language Technologies Department
Vicomtech

How Important are Language Resources for Machine Translation?

Eva Martínez García
emartinez@vicomtech.org

November 20, 2018

Introduction

Machine Translation Systems

- What do they do?

- How do they work?

Language Resources for Machine Translation

Impact on Machine Translation

- Quality of the Data

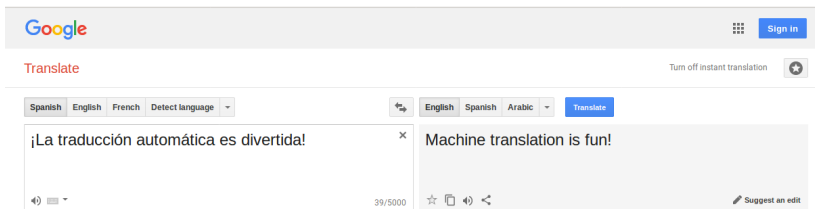
- Amount of Data

- Domain of the Data

Conclusions

Machine Translation is everywhere

Machine Translation is everywhere



Machine Translation is everywhere



Pedro @pmneiva · Oct 31

Trabalha em [#ServiçosTradução](#) na [#AdministraçãoPública](#)? Pretende obter [#MemóriasTradução](#) e outros [#RecursosLinguísticos](#) a partir da sua informação? Assiste à [#ELRIStakeholdersConference](#) em Paris! Mais informação @ [elri-project.eu/StakeholdersCo...](#) [#CEFTelecom](#) [#ELRIProject](#)

ELRI Project @ElriProject

Working in [#PublicAdministration](#) and public [#TranslationServices](#)? Interested in obtaining [#TranslationMemories](#) and other [#LanguageResources](#) from your data? Join us at the [#ELRI Stakeholders Conference](#) in Paris! More info @ [elri-project.eu/StakeholdersCo...](#) ...

 [Translate Tweet](#)



Machine Translation is everywhere



Pedro @pmneiva · Oct 31

Trabalha em [#ServiçosTradução](#) na [#AdministraçãoPública](#)? Pretende obter [#MemóriasTradução](#) e outros [#RecursosLinguísticos](#) a partir da sua informação? Assiste à [#ELRIStakeholdersConference](#) em Paris! Mais informação @ [elri-project.eu/StakeholdersCo...](#) [#CEFTelecom](#) [#ELRIProject](#)

ELRI Project @ElriProject

Working in [#PublicAdministration](#) and public [#TranslationServices](#)? Interested in obtaining [#TranslationMemories](#) and other [#LanguageResources](#) from your data? Join us at the [#ELRI Stakeholders Conference](#) in Paris! More info @ [elri-project.eu/StakeholdersCo...](#) ...

Translated from Portuguese by  Microsoft

Do you work in [#ServiçosTradução](#) na [#AdministraçãoPública](#)? Do you want [#MemóriasTradução](#) to obtain and others [#RecursosLinguísticos](#) from your information? Watch [#ELRIStakeholdersConference](#) Paris! More info @ [elri-project.eu/StakeholdersCo...](#) [#CEFTelecom](#) [#ELRIProject](#)



4



4



Machine Translation is everywhere

... also available for European Public Administrations



- ▶ Translate plain texts or formatted documents
- ▶ State of the Art MT Systems
- ▶ 24 official EU languages, Icelandic and Norwegian
- ▶ Stand-alone or Integrated service
- ▶ EU institutions/agencies or national public administrations in EU countries, Norway and Iceland.

<https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/eTranslation>

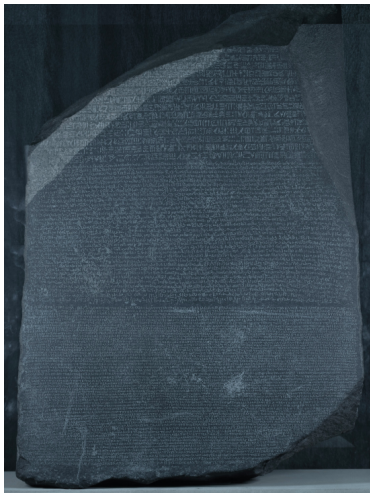
Machine Translation Systems

What does an MT System do?

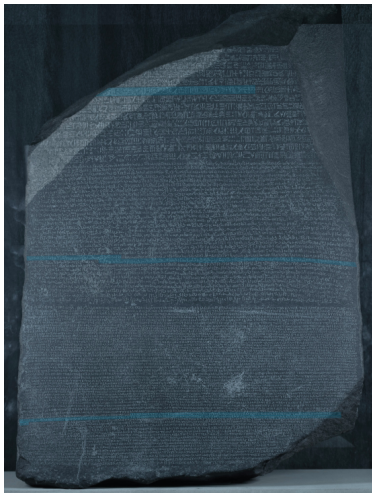


- ▶ Typically translates sentence by sentence
- ▶ Typically uses Machine Learning models to perform translation

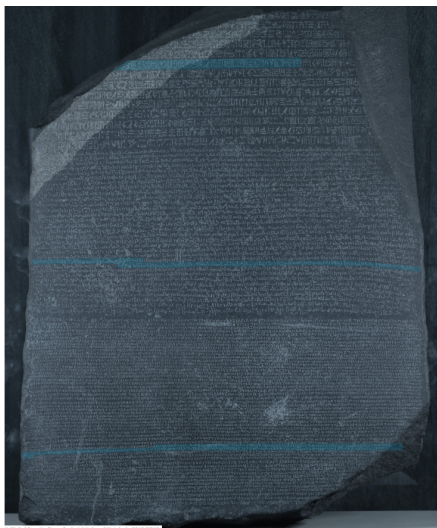
MT Systems learn from Multilingual Data



MT Systems learn from Multilingual Data



MT Systems learn from Multilingual Data



mtwsw t'hy [nry n nsw-bj] (Ptlwmy's nry,n) d.t mxy Pthj ntr pn nb nfr.w k' tw mnt (Ptlwmy's) nry,n) S'q.t wh' n'f pr
(Ptlwmy's) n[ht(n)] [Km.t] [...]

mtwsw dd naf (Ptlwmy's) nq Bay nti w p'nf whm (Ptlwmy's) iir nh.t Kmy sm w' twtw (n) p' ntr (n) t' mwt swsf d.t
naf hpt (n) qny n p' ppy nry sp-2 (n) p' m' nti wnt n p' ppy nsw-w ir t-b.t wp.t (n) mnt-Kmy

η προσονομασθησα (Πτολεμαίω) του επιμενάντος της Αιγυπτου, η παρορτηζεται ο κυριωτατος θεός του ιερου,
διδους αυτων εθλον νηπτεων, ε εστιν κατασκευασμένη των των Αιγυπτιαων

MT Systems learn from Aligned Data

Cuando yo tenía seis años vi en un *libro* sobre la selva virgen que se titulaba "Historias vividas", una magnífica lámina. Representaba una serpiente boa que se tragaba a una fiera. Esta es la copia del dibujo.

Once when I was six years old I saw a magnificent picture in a *book*, called True Stories from Nature. It was a picture of a boa constrictor in the act of swallowing an animal. Here is a copy of the drawing.



MT Systems learn from Aligned Data

Cuando yo tenía seis años vi en un **libro** sobre la selva virgen que se titulaba "Historias vividas", una magnífica lámina. Representaba una serpiente boa que se tragaba a una fiera. Esta es la copia del **dibujo**.

Once when I was six years old I saw a magnificent picture in a **book**, called True Stories from Nature. It was a picture of a boa constrictor in the act of swallowing an animal. Here is a copy of the **drawing**.



MT Systems learn from Aligned Data

Cuando yo tenía seis años vi en un **libro** sobre la selva virgen que se titulaba "Historias vividas", una magnífica lámina. Representaba una serpiente boa que se tragaba a una fiera. Esta es la copia del **dibujo**.

En el **libro** se afirmaba: "La serpiente boa se traga su presa entera, sin masticarla.

Luego ya no puede moverse y duerme durante los seis meses que dura su digestión".



Once when I was six years old I saw a magnificent picture in a **book**, called True Stories from Nature. It was a picture of a boa constrictor in the act of swallowing an animal. Here is a copy of the **drawing**.

In the **book** it said: "Boa constrictors swallow their prey whole, without chewing it.

After that they are not able to move, and they sleep through the six months that they need for digestion."

MT Systems learn from Aligned Data

Cuando yo tenía seis años vi en un **libro** sobre la selva virgen que se titulaba "Historias vividas", una magnífica lámina. Representaba una serpiente boa que se tragaba a una fiera. Esta es la copia del **dibujo**.

En el **libro** se afirmaba: "La serpiente boa se traga su presa entera, sin masticarla.

Luego ya no puede moverse y duerme durante los seis meses que dura su digestión".

Reflexioné mucho en ese momento sobre las aventuras de la jungla y a mi vez logré trazar con un lápiz de colores mi primer **dibujo**. Mi **Dibujo** Número 1 era de esta manera:



Once when I was six years old I saw a magnificent picture in a **book**, called True Stories from Nature. It was a picture of a boa constrictor in the act of swallowing an animal. Here is a copy of the **drawing**.

In the **book** it said: "Boa constrictors swallow their prey whole, without chewing it.

After that they are not able to move, and they sleep through the six months that they need for digestion."

I pondered deeply, then, over the adventures of the jungle. And after some work with a colored pencil I succeeded in making my first **drawing**. My **Drawing** Number One. It looked like this:

MT Systems learn from Aligned Data

Cuando yo tenía **seis años** vi en un **libro** sobre la selva virgen que se titulaba "Historias vividas", una magnífica lámina. Representaba una serpiente boa que se tragaba a una fiera. Esta es la copia del **dibujo**.

En el **libro** se afirmaba: "La serpiente boa se traga su presa entera, sin masticarla.

Luego ya no puede moverse y duerme durante los **seis meses** que dura su digestión".

Reflexioné mucho en ese momento sobre las aventuras de la **jungla** y a mi vez logré trazar con un **lápiz** de colores mi primer **dibujo**.

Mi **Dibujo** Número 1 era de esta manera:



Once when I was **six years** old I saw a magnificent picture in a **book**, called True Stories from Nature. It was a picture of a boa constrictor in the act of swallowing an animal. Here is a copy of the **drawing**.

In the **book** it said: "Boa constrictors swallow their prey whole, without chewing it.

After that they are not able to move, and they sleep through the **six months** that they need for digestion."

I pondered deeply, then, over the adventures of the **jungle**. And after some work with a colored **pencil** I succeeded in making my first **drawing**. My **Drawing** Number One. It looked like this:

MT Systems learn from Aligned Data

Cuando yo tenía **seis años** vi en un **libro** sobre la selva virgen que se titulaba "Historias vividas", una **magnífica** lámina. Representaba una **serpiente boa** que se tragaba a una fiera. Esta es la copia del **dibujo**.

En el **libro** se afirmaba: "La **serpiente boa** se traga su presa entera, sin masticarla.

Luego ya no puede moverse y duerme durante los **seis meses** que dura su digestión".

Reflexioné mucho en ese momento sobre las **aventuras** de la **jungla** y a mi vez logré trazar con un **lápiz** de colores mi primer **dibujo**.

Mi **Dibujo** Número 1 era de esta manera:



Once when I was **six years** old I saw a **magnificent** picture in a **book**, called True Stories from Nature. It was a picture of a **boa constrictor** in the act of swallowing an animal. Here is a copy of the **drawing**.

In the **book** it said: "**Boa constrictors** swallow their prey whole, without chewing it. After that they are not able to move, and they sleep through the **six months** that they need for digestion."

I pondered deeply, then, over the **adventures** of the **jungle**. And after some work with a colored **pencil** I succeeded in making my first **drawing**. My **Drawing** Number One. It looked like this:

MT systems need data!

MT systems need **LOTS** of data!

What kind of data do we need to build MT systems?

What kind of data do we need to build MT systems?

- ▶ **Aligned** data in different languages to learn how to translate

What kind of data do we need to build MT systems?

- ▶ **Aligned** data in different languages to learn how to translate
- ▶ **Clean** data: misspellings, language, codification

What kind of data do we need to build MT systems?

- ▶ **Aligned** data in different languages to learn how to translate
- ▶ **Clean** data: misspellings, language, codification
- ▶ **Generic** data to cover as much language as possible

What kind of data do we need to build MT systems?

- ▶ **Aligned** data in different languages to learn how to translate
- ▶ **Clean** data: misspellings, language, codification
- ▶ **Generic** data to cover as much language as possible
- ▶ **Domain-specific** data to produce more accurate translations

MT systems learn to generate from the data they see.

How much data do we need?

How much data do we need?

- **Generic** Coverage: ideally millions of sentence pairs

Corpora English-Spanish

Corpus	# sentences (app.)	# words (app.)
Europarl	2 M	50M
MultiUN	17M	440M

How much data do we need?

- ▶ **Generic** Coverage: ideally millions of sentence pairs

Corpora English-Spanish		
Corpus	# sentences (app.)	# words (app.)
Europarl	2 M	50M
MultiUN	17M	440M

- ▶ **Domain** Coverage:
 - ▶ the **more**, the **better**, but
 - ▶ **thousands** of sentence pairs are already helpful

POOR data results in POOR translations

Source	
In their filing, they referred to the Supreme Court's 1966 decision overturning Dr. Sheppard's conviction because of the trial coverage and made it clear they would claim that the new media required new rules.	

POOR data results in POOR translations

Source	Noisy MT
In their filing, they referred to the Supreme Court's 1966 decision overturning Dr. Sheppard's <i>conviction</i> because of the trial coverage and made it clear they would claim that the new media required new rules.	En su texto, refirieron a la 1966 decisión del Supremo Tribunal del doctor doctor Tribunal Supremo Supremo .

POOR data results in POOR translations

Source	Noisy MT
In their filing, they referred to the Supreme Court's 1966 decision overturning Dr. Sheppard's <i>conviction</i> because of the trial coverage and made it clear they would claim that the new media required new rules.	En su texto, refirieron a la 1966 decisión del Supremo Tribunal del doctor doctor Tribunal Supremo Supremo de la <i>convicción</i> del Sheppard sobre los medios.

POOR data results in POOR translations

Source	Noisy MT
In their filing, they referred to the Supreme Court's 1966 decision overturning Dr. Sheppard's <i>conviction</i> because of the trial coverage and made it clear they would claim that the new media required new rules.	En su texto, refirieron a la 1966 decisión del Supremo Tribunal del doctoer Sheppard doctoer Tribunal Supremo Supremo de la <i>convicción</i> del Sheppard sobre los medios.

POOR data results in POOR translations

Source	Noisy MT
In their <i>filing</i> , they referred to the <i>Supreme</i> Court's 1966 decision overturning <i>Dr.</i> Sheppard's <i>conviction</i> because of the trial coverage and made it clear they would claim that the new media required new rules.	En su <i>filing</i> , refirieron a la <i>Supreme</i> 1966 Corte decisión de Sheppard el <i>docteur conviction</i> porque la prueba de <i>couverture und machte es klar</i> que clamarían <i>Neue Regeln</i> para el nuevo media.

Impact on Machine Translation

Amount

MORE data leads to BETTER translations

Source	Low Resource MT
In their filing , they referred to the Supreme Court's 1966 decision overturning Dr. Sheppard's conviction because of the trial coverage and made it clear they would claim that the new media required new rules.	En su <UNK> , refirieron a la <UNK> 1966 <UNK> decisión de <UNK> el doctor <UNK> convicción porque la prueba de cubierta y hicieron claro que clamarían que el nuevo medio requería nuevas reglas.

MORE data leads to BETTER translations

Source	Low Resource MT
In their filing , they referred to the Supreme Court's 1966 decision overturning Dr. Sheppard's conviction because of the trial coverage and made it clear they would claim that the new media required new rules.	En su <UNK> , refirieron a la <UNK> 1966 <UNK> decisión de <UNK> el doctor <UNK> convicción porque la prueba de cubierta y hicieron claro que clamarían que el nuevo medio requería nuevas reglas.

Impact on Machine Translation

Amount

MORE data leads to BETTER translations

Source	Low Resource MT
In their filing , they referred to the Supreme Court's 1966 decision overturning Dr. Sheppard's conviction because of the trial coverage and made it clear they would claim that the new media required new rules.	En su <UNK> , refirieron a la <UNK> 1966 <UNK> decisión de <UNK> el doctor <UNK> convicción porque la prueba de cubierta y hicieron claro que clamarían que el nuevo medio requería nuevas reglas.

MORE data leads to BETTER translations

Source	Low Resource MT
In their filing , they referred to the Supreme Court's 1966 decision overturning Dr. Sheppard's conviction because of the trial coverage and made it clear they would claim that the new media required new rules.	En su filing , refirieron a la Supreme 1966 Court decisión de Sheppard el doctor overturning convicción porque la prueba de cubierta y hicieron claro que clamarían que el nuevo medio requería nuevas reglas.

PRECISE translations need data from the DOMAIN

Source	General MT
In their filing , they referred to the Supreme Court's 1966 decision overturning Dr. Sheppard's conviction because of the trial coverage and made it clear they would claim that the new media required new rules.	En su limadura , refirieron a la decisión de 1966 del Tribunal Supremo que volcaban la convicción del Dr. Sheppard debido a la cobertura de ensayo y clarificado demandarían que los nuevos medios requirieron nuevas reglas.

PRECISE translations need data from the DOMAIN

Source	Adapted MT
In their filing , they referred to the Supreme Court's 1966 decision overturning Dr. Sheppard's conviction because of the trial coverage and made it clear they would claim that the new media required new rules.	En su presentación , se refirieron a la decisión del Tribunal Supremo de 1966 que revocó la condena del Dr. Sheppard debido a la cobertura del juicio y dejaron claro que afirmarían que los nuevos medios requerían nuevas reglas.

MT is useful!

MT is useful!

What do we need to have a good MT system?

MT is useful!

What do we need to have a good MT system?

- ▶ LOTS of data

MT is useful!

What do we need to have a good MT system?

- ▶ LOTS of data
- ▶ QUALITY data

MT is useful!

What do we need to have a good MT system?

- ▶ LOTS of data
- ▶ QUALITY data
- ▶ DOMAIN Specific data

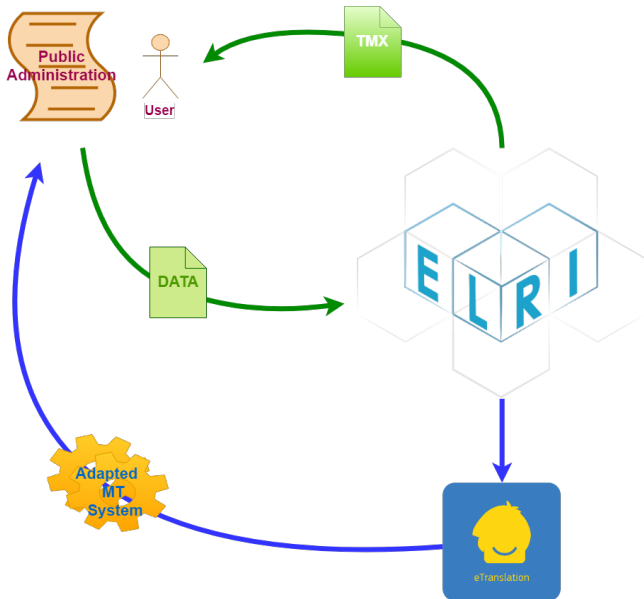
MT is useful!

What do we need to have a good MT system?

- ▶ LOTS of data
- ▶ QUALITY data
- ▶ DOMAIN Specific data

It is important to SHARE language resources!

How can you help improve MT quality?





E

Thank You for Your Attention!

L

R

I