English-Catalan Neural Machine Translation in the Biomedical Domain through the cascade approach

Marta R. Costa-jussà
marta.ruiz@upc.edu

Noe Casas
contact@noecasas.com

Maite Melero
maite.melero@upf.edu

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

AGAUR Agència de Gestió d'Ajuts Universitaris i de Recerca

MINISTERIO DE ECONOMÍA, INDUSTRIA Y COMPETITIVIDAD

Plan de Impulso de las Tecnologías del Lenguaje

Motivation for English-Catalan translation for biomedical domain

- English has become *de facto* universal language for scientific communication.

- Catalonia has become a hub of global biomedical research.

- Also useful for patient-physician communication, in a country with over 2,5 M visitors per year.

- Only an estimated 30% of the population has good competence of English.

Catalan language

- Western romance language.
- Latin alphabet + 2 written accents (`´) + diaeresis + ç + mid dot (·)
- Co-official language in Spanish territories of Catalonia, Balearic Islands and Valencia. Official language in Andorra.
- 4.1 million native speakers + 5 million second-language speakers.

English-Catalan translation for biomedical domain

- Neural MT Technology
- Low (or zero) resource language pair for the domain.
- Use a third language (Spanish) as pivot language.
- English-Spanish corpus from biomedical domain.
- Spanish-Catalan corpus from a diferent domain (news).

Datasets

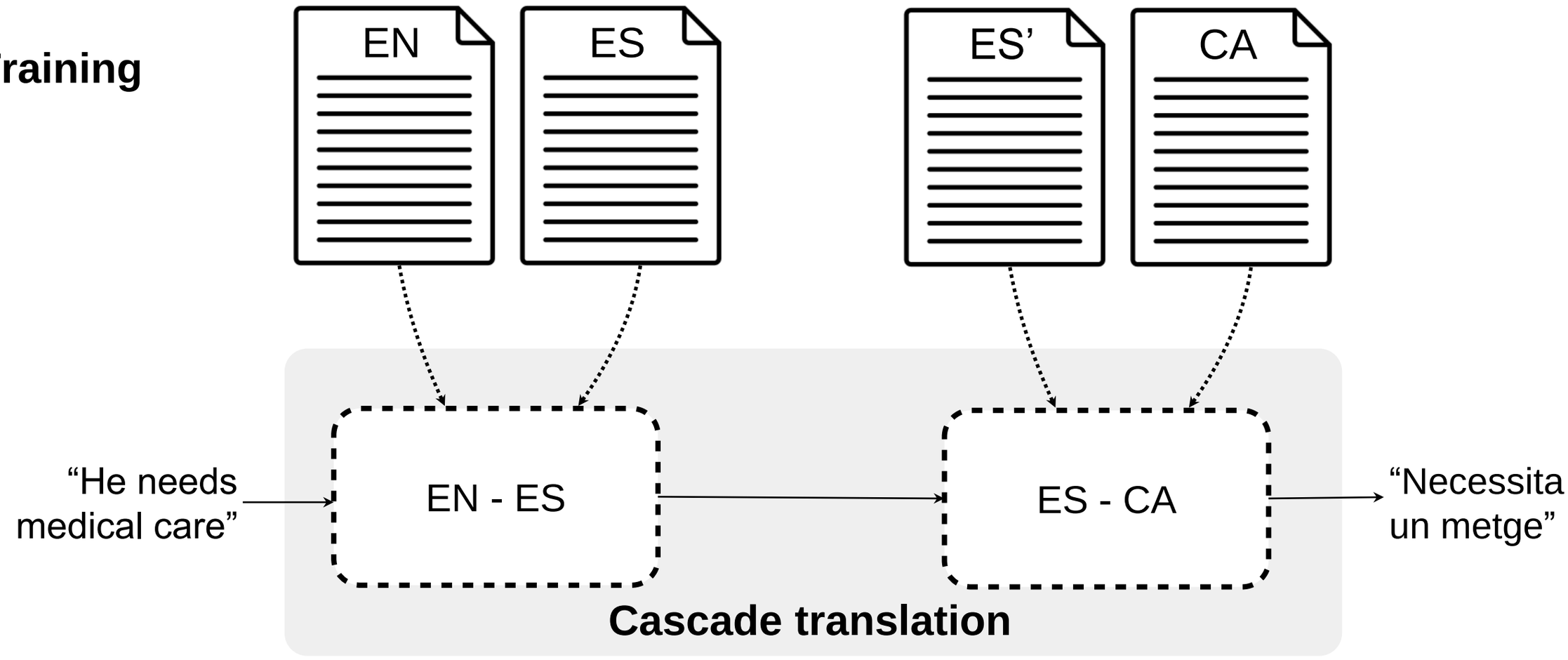
Table 1: Size of the parallel training corpora

Language Pair	**Corpus**	Language	Segments	Words	Vocab
En-Es	Biomedical	En	$0.9 \cdot 10^6$	$20.5 \cdot 10^6$	$296 \cdot 10^3$
		Es		$21.9 \cdot 10^6$	$309 \cdot 10^3$
Es-Ca	El Periódico	Es	$6.5 \cdot 10^6$	$165.1 \cdot 10^6$	$736 \cdot 10^3$
		Ca		$178.9 \cdot 10^6$	$713 \cdot 10^3$

Table 2: Size of the test set

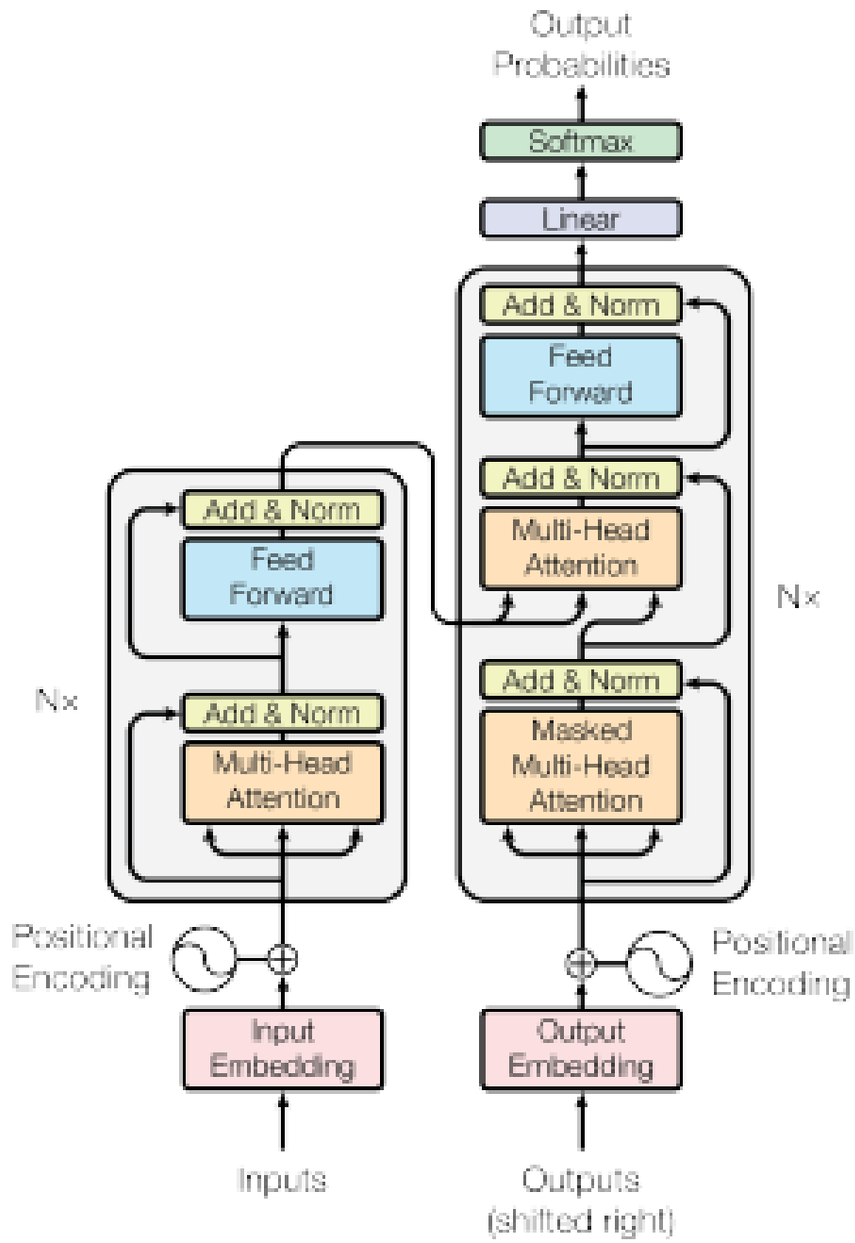
Language Pair	**Corpus**	Language	Segments	Words	Vocab
En-Es	Biomedical	En	1000	$26.1 \cdot 10^3$	$6.1 \cdot 10^3$
		Es		$27.4 \cdot 10^3$	$6.6 \cdot 10^3$
Es-Ca	El Periódico	Es	2244	$56.0 \cdot 10^3$	$12.2 \cdot 10^3$
		Ca		$60.7 \cdot 10^3$	$11.7 \cdot 10^3$

Cascade translation



Neural Architecture: Transformer

- Based on a feed-forward encoder-decoder scheme based solely on attention mechanisms.
- No recurrence or convolutions, as proposed in initial architectures.
- Multi-head attention allows to train several attention modules in parallel (less time).
- Self-attention allows use of same sentence as input (best results on long-distance, co-reference, etc.)



Attention is all you need
Vaswani et al., 2017

Pre-processing of the data

- Tokenization (including separation of clitics and contractions (EN are+n’t, ES hacer+lo, de+el) for EN-ES) (Freeling pipeline)

- Sentence segmentation of max 50 words.

- Problem: out-of-vocabulary words, esp. inflected verbs, domain-specific technical terms => Subword mechanism

Sub-word analysis

1) take all the words present in the vocabulary along with their frequency of occurrence,

2) extract all their characters,

3) use those characters as initial token vocabulary,

4) iteratively modify the vocabulary by joining tokens from the current vocabulary whose combination appears more frequently in the training corpus. Repeat until desired vocabulary size is achieved (c. 32K).

Results

Language	System	BLEU
EN2ES	Direct	46.55
ES2CA	Direct	86.89
EN2CA	Cascade	41.38

EN2CA test set: English source of EN-ES biomedical test set, with manual Catalan translations

More natural translation (compared to rule-based)

	Original	NMT	Rule-based ES-CA
Insertion of a *weak* pronoun	*including implementation of monitoring programs*	*incloent-hi la implementació de programes de vigilància*	*incloent la implementació de programes de vigilància*
Better choice of preposition	*at Samanga paramo a total of 67 species were registered*	*a l ' erm de Samanga , s ' hi van registrar 67 espècies*	*en l ' erm de Samanga , es van registrar 67 espécies*
Better lexical selection	*genomic fingerprint*	*empremta genòmica*	*petjada genòmica*
Phrase re-ordering	*the most important Swiss medical journal of the 20th century*	*la revista mèdica suïssa més important del segle XX*	*la més important revista mèdica suïssa del segle XX*

Annoying typical NMT errors

	Original	NMT	Corrected Translation
NMT extra verbosity	*differ in their permeabilizing activity*	*diferències en la permeabilització ació ació ació ació i la permeabilització*	*diferències en la permeabilització*
NMT drops text	*the micronucleus induction test in mouse bone marrow that determines clastogenic and aneugenic damage*	*l ' assaig d ' induc*	*l'assaig d'inducció de micronuclis en medul·la òssia de ratolí, que determina dany clastogènic i aneugènic*
NMT inserts text out of the blue	*one hundred and forty-one fulfilled the inclusion criteria*	*cent quaranta anys i una complien els criteris d'inclusió*	*cent quaranta-una complien els criteris d'inclusió*

Errors derived from the cascade approach

	Original	EN-ES	ES - CA
Propagation of errors occuring in the first step of translation	*infected pork*	*carne de cerdo infectada con cisticercos*	*carn de porc infectada amb cisticercs*
Misadjustments due to training with different varieties of pivot	*four patient died .*	*los cuatro pacientes restantes se encuentran fallecidos* (LAm)	*els quatre pacients restants es troben morts*
Pivot introduces new ambiguity	*calcium oxalate crystal formation*	*cristales de oxalato de calcio*	*vidres (en. glass) d ' oxalat de calcio (cristalls)*

Inspection of sub-word mechanism

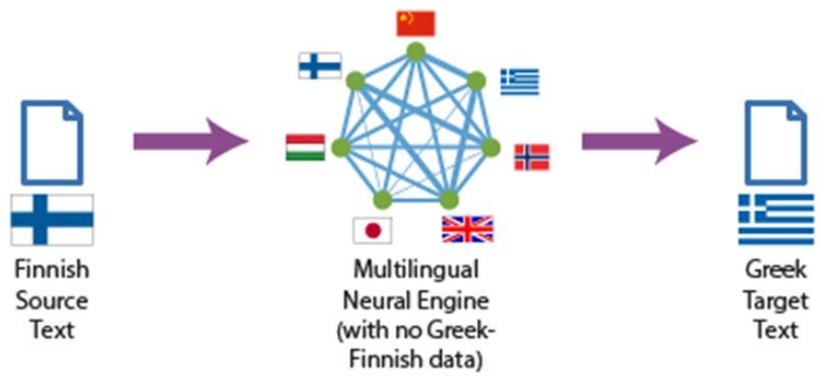
	Original	NMT	Correct translations
Successful translation of OOV words from biomedical domain	*Bioassays* *Physicochemical* *Cryopreservation* *Toxicogenetic* *Pleomorphism*	*Bioassajos* *Fisicoquímics* *Crioconservació* *Toxicogenètic* *Pleomorfisme*	
Good try but ... no	*Metacestode* *Peptide* *Survival* *Physician-teacher* *Undetermined* *Endemism* *Prelarviposition* *Subclavian* *Brief psychotherapy*	*Metacestot* *Pèptits* *Sobreviuència* *Medicerodocent* *Indeacabat* *Endemateixos* *Prelatrvsició* *Subcràpia* *Psicobreu teràpia*	*Metacestode* *Pèptids* *Supervivència* *Metge-docent* *Indeterminat* *Endemismes* *Prelarviposició* *Subclàvia* *Psicoteràpia breu*

Conclusions

- We have described the data resources and architecture to build an English-Catalan neural MT system, in the biomedical domain,
- providing state-of-the-art results,
- without need of any EN-CA parallel resources,
- using the latest architectures in NMT, based on attention mechanisms,
- and one standard pivot architecture, namely, the cascade model.

Future lines of research

- Take results as baseline and compare with other pivot approaches:
  - Pseudo-corpus pivot translation, i.e. generate a new CA biomedical corpus using ES-CA system.
  - Zero-shot translation (existing parallel corpora concatenated as one single training corpus): one pivot or multiple pivots



- Unsupervised translation based on cross-lingual word embeddings (using only mono-lingual corpora)

Gràcies per la vostra atenció!

感谢您的关注

Merci pour votre attention!

あなたの注意のおかげで

Danke für Ihre Aufmerksamkeit!

شكرا على انتباهك

Eskerrak zuen arretagatik!

¡Gracias por vuestra atención!