Journey through Natural Language Processing

Cristina España-Bonet

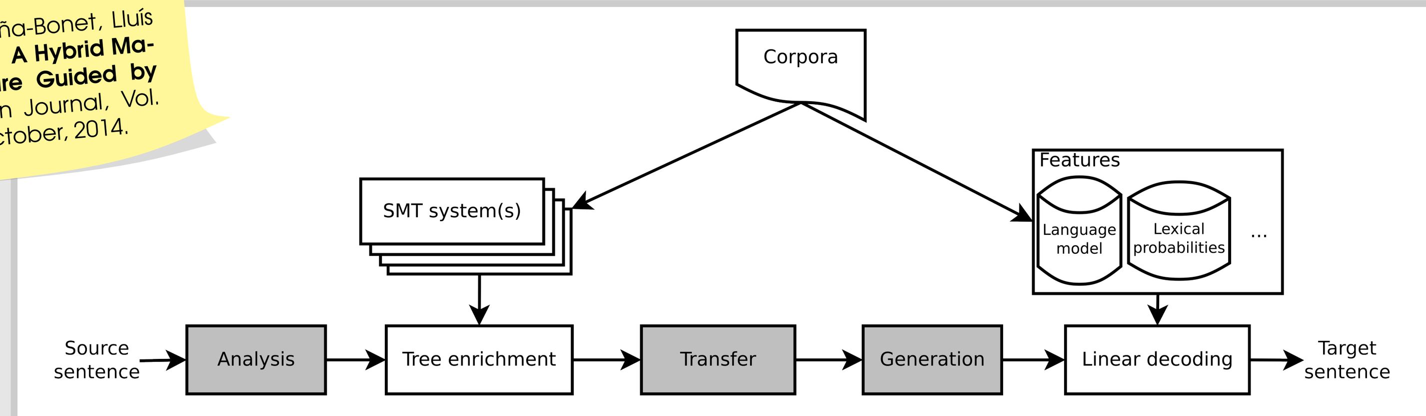
Universitat Politècnica de Catalunya – BarcelonaTech

4Google NLP SUMMIT 2015

Hybrid Machine Translation

SMatxinT, a HMT Architecture Guided by Syntax

Gorka Labaka, Cristina España-Bonet, Lluís Màrquez and Kepa Sarasola. **A Hybrid Machine Translation Architecture Guided by Syntax.** Machine Translation Journal. Vol. 28, Issue 2, pages 91–125, October, 2014.

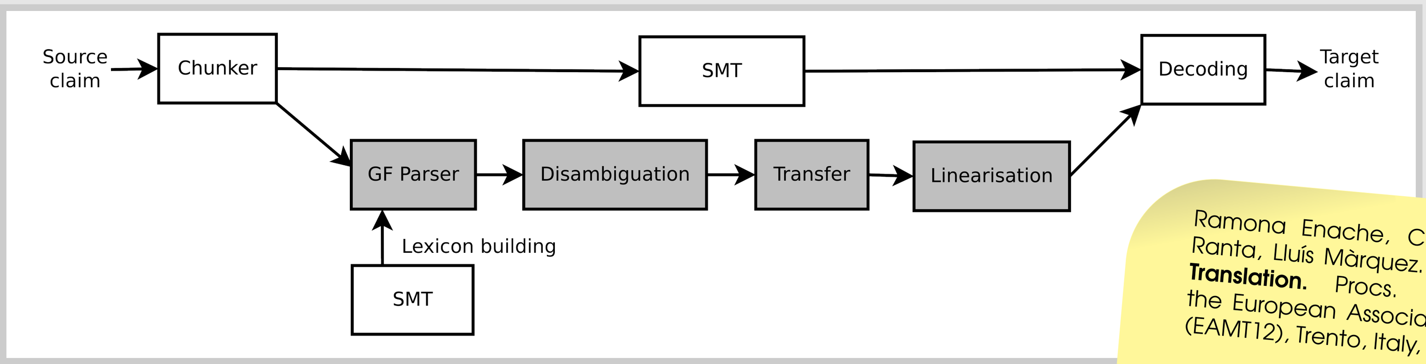


	TER	BLEU	NIST	GTM-2	MTR-st	RG-S*	ULC
Matxin	73.57	14.29	6.05	22.62	20.27	15.90	41.02
SMT_b	68.49	15.93	6.45	23.64	21.59	16.39	44.56
SMT_g	68.93	15.21	6.43	23.44	21.84	17.66	44.20
SMatxinT (m)	**66.70**	**17.14**	**6.72**	**24.58**	**22.52**	**18.66**	**48.00**
SMatxinT (r)	**66.63**	**17.18**	**6.73**	**24.59**	**22.51**	**18.52**	**48.03**

Es2Eu

- Candidate chunk translations are calculated with an SMT system and used to enrich the RBMT tree-based source representation with more alternatives
- The most probable combination among the available fragments is obtained with a monotone statistical decoding (m) following the order provided by the RBMT system
- The evaluation on news (out-of-domain) shows a preference for SMatxinT. There is no necessity for reordering the RBMT choice (r)

MOLTO, a HMT for Patent Translation



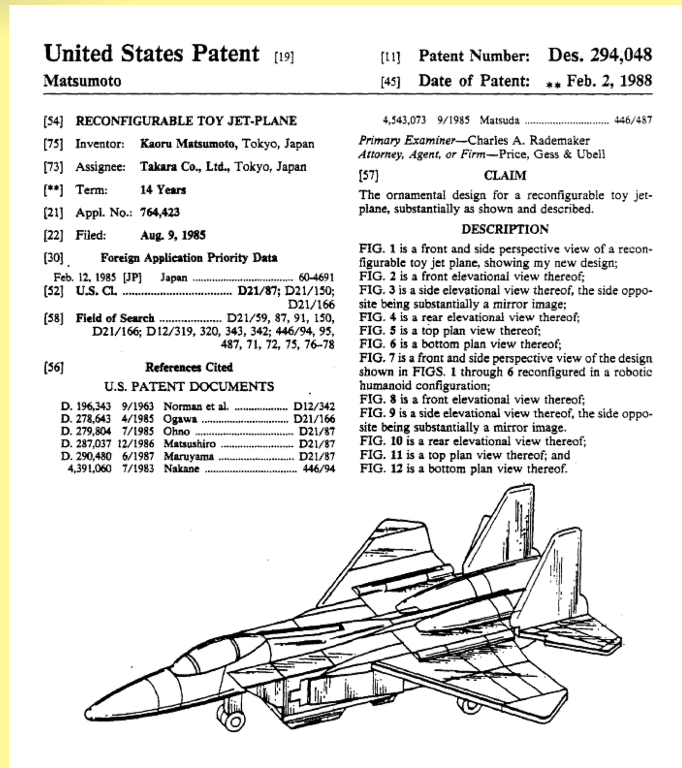
Ramona Enache, Cristina España-Bonet, Aarne Ranta, Lluís Màrquez. **A Hybrid System for Patent Translation.** Procs. 16th Annual Conference of the European Association for Machine Translation (EAMT12), Trento, Italy, May 8-30, 2012.

- A GF translator (RBMT) is built for the specific domain which, in turn, uses an in-domain SMT system to build its lexicon
- Another SMT system is on top of the GF to translate those phrases not covered by the grammar. Hard integration (HI) vs. Soft integration (SI)
- Evaluations consistently show a preference for the SI hybrid system in front of the two individual translators

	TER	BLEU	NIST	GTM-2	MTR-p	RG-S*	ULC
GF	58.90	26.56	5.57	22.74	38.76	29.00	16.17
SMT	25.32	63.18	9.99	44.58	71.64	72.65	67.14
HI	31.24	55.88	9.24	38.81	67.30	67.80	58.84
SI1.0	25.10	63.56	10.02	**44.86**	**71.96**	72.89	67.56
SI0.5	**25.02**	**63.60**	**10.03**	44.84	71.94	**72.93**	**67.60**
SI0.0	25.36	63.15	9.99	44.54	71.60	72.66	67.11

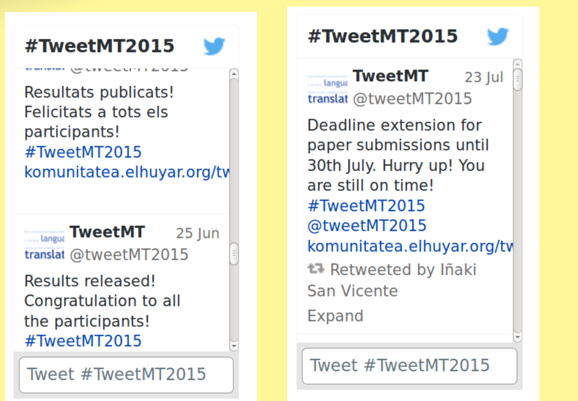
En2Fr

Piece of news



Patent

Wikipedia article



Tweets

Document-Level Machine Translation

- Using a document-level (DL) decoder and considering a feature function that rewards coherent translations
- Capturing semantic information in corpora with bilingual word vector models that can be integrated as semantic space language models (biSSM) at translation time

	TER	BLEU	NIST	MTR-p	RG-S*	SP-Op	ULC
DL	72.83	28.33	7.46	23.22	30.36	19.38	77.14
DL + monoSSM	72.61	28.48	7.52	23.28	30.33	19.61	77.49
DL + biSSM	**72.56**	**28.58**	**7.66**	**23.31**	**30.38**	**19.78**	**77.89**

En2Es on News

Eva Martínez Garcia, Cristina España-Bonet, Lluís Màrquez. **Document-Level Machine Translation with Word Vector Models.** Procs. 18th Annual Conference of the European Association for Machine Translation (EAMT15), pages 59-66, Antalya, Turkey, May 2015.

Eva Martínez Garcia, Cristina España-Bonet, Lluís Màrquez. **The UPC TweetMT participation: Translating Formal Tweets using Context Information.** Procs. "XXXI Congreso de la Sociedad Española de Procesamiento de lenguaje natural", Alacant, Spain, September 2015.

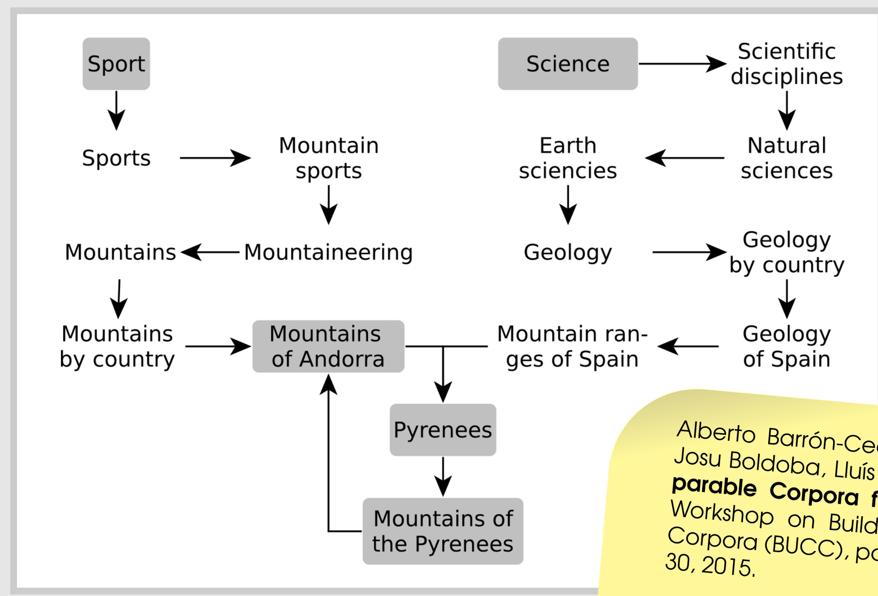
	TER	BLEU	NIST	GTM-2	MTR-e	RG-S*	Ol	ULC
SMT	13.56	78.06	12.04	**73.02**	54.09	82.21	83.31	66.62
DL	**13.44**	**78.19**	**12.07**	72.97	**54.15**	**82.45**	**83.48**	**67.11**

Es2Ca on Tweets

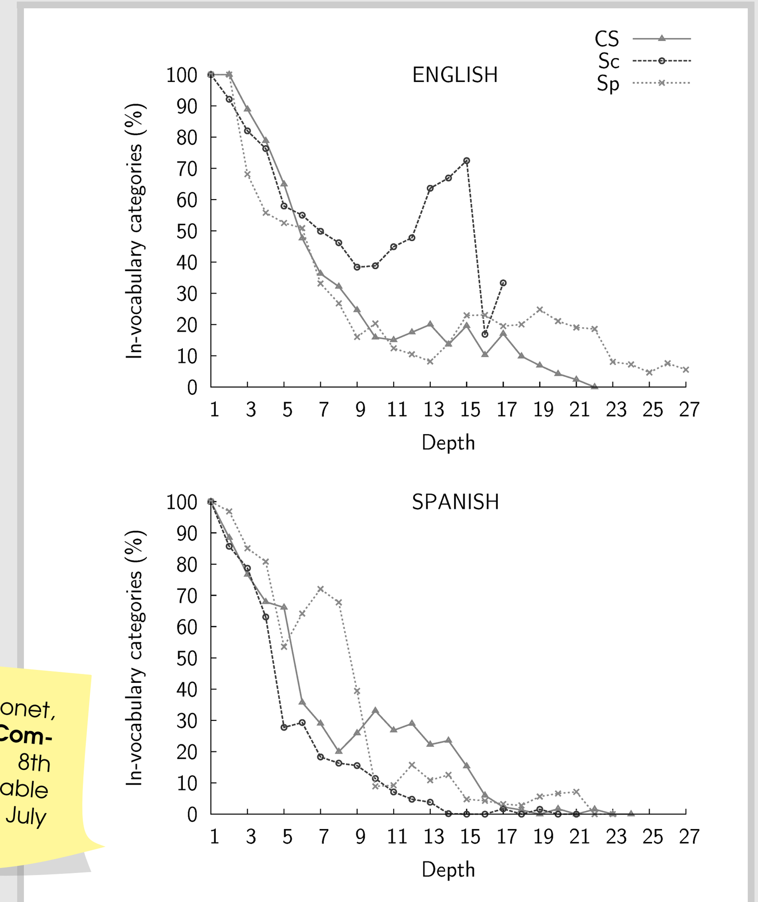
Comparable Corpora from Wikipedia

Domain Corpora

Given a domain, explore Wikipedia's category graph as a tree —breadth-first search departing from the main category visiting nodes only once



Alberto Barrón-Cedeño, Cristina España-Bonet, Josu Boldoba, Lluís Màrquez. **A Factory of Comparable Corpora from Wikipedia.** Procs. 8th Workshop on Building and Using Comparable Corpora (BUCC), pages 3–13, Beijing, China, July 30, 2015.



- Define an in-domain vocabulary (MFTs in *root* articles)
- Stopping criterion: levels with a mimimun percentage of categories with in-domain vocabulary in the title

	Articles		Depth			
	50%	60%	50%		60%	
	en-es	en-es	en	es	en	es
CS	18,168	8,251	6	5	5	5
Sc	161,130	21,459	6	4	4	4
Sp	72,315	1,980	8	8	3	4

CS: Computer Science, Sc: Science, Sp: Sports

Comparable Corpora

- Articles are related across languages via *interlanguage links*
- In-domain CC: monolingual extracted, union or intesection
- Usefulness of the corpora confirmed for training SMT systems

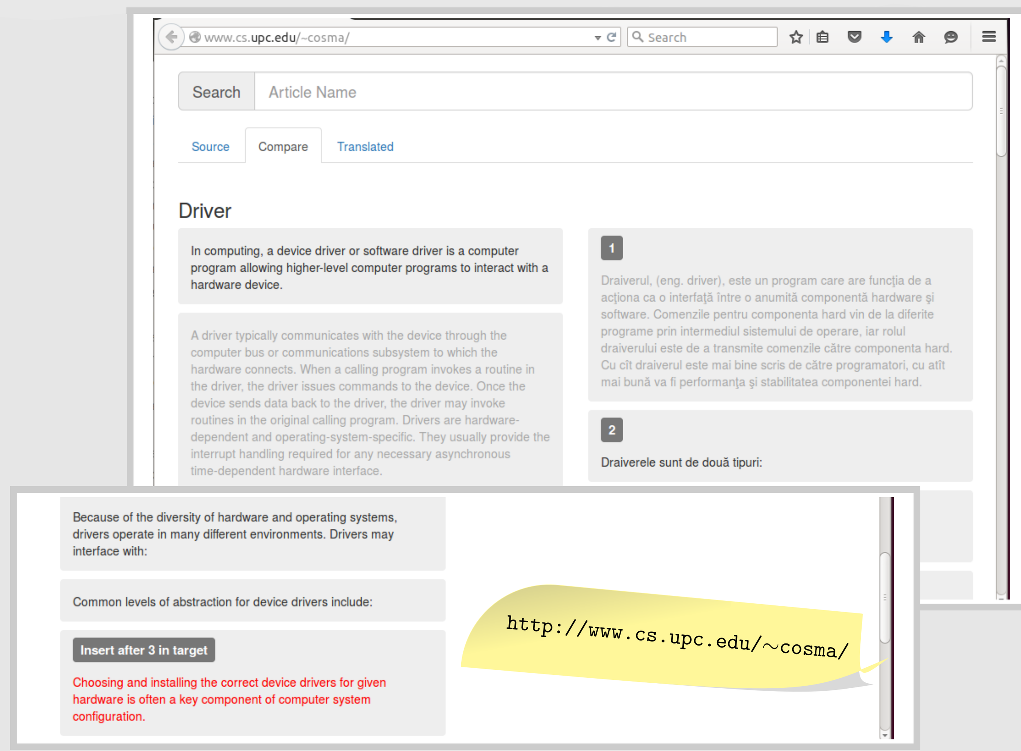
Wikiparable, CC Extraction (ongoing work)

	Graph-based			IR-based	
	#Arts/Cat	Depth	ρ_{Spear}	#Arts/Cat	ρ_{Spear}
1 English	50,514±121,881	5.9±2.8	0.38±0.20	61,239±73,248	0.27±0.15
3 French	8,278±26,483	4.3±1.9	0.39±0.23	18,158±17,871	0.15±0.13
4 Spanish	6,638±17,050	4.4±2.1	0.38±0.22	21,490±19,605	0.18±0.14
2 German	2,752±9,573	3.4±1.9	0.42±0.25	12,887±18,876	0.13±0.13
6 Arabic	2,999±9,546	3.6±2.3	0.45±0.28	4,622±4,082	0.12±0.13
7 Romanian	1,398±8,683	3.4±1.8	0.42±0.26	1,750±1,839	0.06±0.13
5 Catalan	1,140±4,693	3.3±1.9	0.44±0.22	5,959±5,058	0.14±0.13
8 Basque	440±1,654	3.1±1.5	0.48±0.25	1,819±2,732	0.13±0.13
9 Greek	356±1,982	2.8±1.6	0.55±0.22	1,604±2,378	0.16±0.16
10 Occitan	104±598	2.4±1.3	0.51±0.29	419±2,040	0.14±0.17

Mean values over 743 categories — Results for the most restrictive model for each architecture

- 700+ main categories (domains) and 10 languages analised
- Graph-based and IR-based architectures compared
- For each architecture, parameters can be varied to obtain larger (more general) or smaller (more concrete) corpora
- Spearman correlation with an in-domain corpus (root articles) or the density of in-domain terms can be used to evaluate the degree of *in-domainess*
- The tool written in Java will be publicly available for both methods of CC extraction

Wikipedia Enrichment



http://www.cs.upc.edu/~cosma/

Prototype for including relevant information into a Wikipedia article in language L_1 from the same article in L_2

1. Paragraph alignment (similarity measures)
2. Relevant paragraph identification (from unaligned paragraphs)
3. Determination of the insertion position into the target
4. Preliminar translation of the relevant paragraphs by a translation engine