Comparing Rule-Based and Data-Driven approaches to Spanish-to-Basque Machine Translation

Gorka Labaka[1], Nicolas Stroppa[2], Andy Way[2], Kepa Sarasola[1]

1 - IXA NLP group
University of the Basque Country
{jiblaing, kepa.sarasola}@ehu.es

2 - National Centre for Language Technology
Dublin City University
{nstroppa, away}@computing.dcu.ie

BASQUE LANGUAGE

- Basque is a **morphologically-rich** agglutinative language.
  Translating to Basque requires a huge generation of morphological information.
  Some Spanish words, like prepositions or articles, correspond to Basque suffixes. In case of ellipsis more than one suffix can be added to the same word.

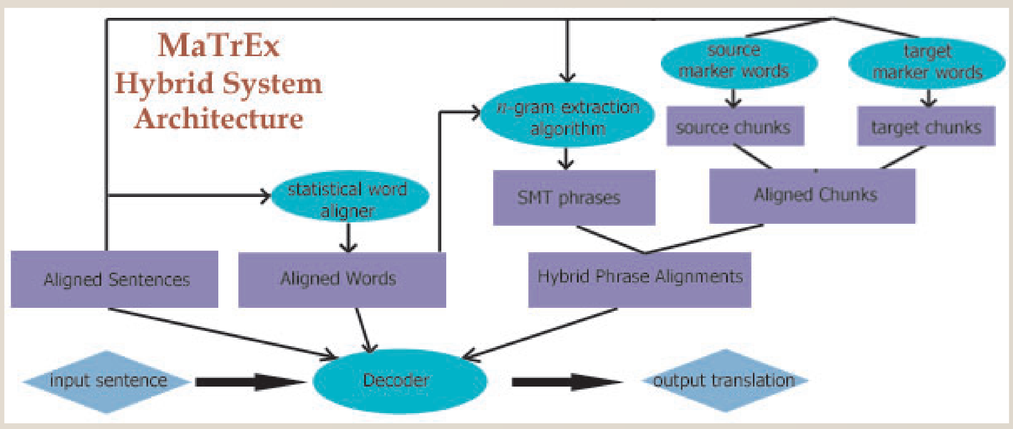
etxe	etxeko	etxekoa	etxekoarengana
/house/	*/of the house/*	*/the one of the house/*	*/towards the one of the house/*

- Basque is **free-order**: the order of the constituents into the sentence is free.
  Both, "Gizona etxera doa" and "Etxera doa gizona" (*/the man goes home/*), are correct in Basque.
- Basque is a **low-density language**:
  - Basque is spoken by almost 1,000,000 people, but only for 500,000 people it is their first language.
  - There are few corpora available and they are smaller than corpora available for other languages.

MACHINE TRANSLATION SYSTEMS

MATREX SYSTEM

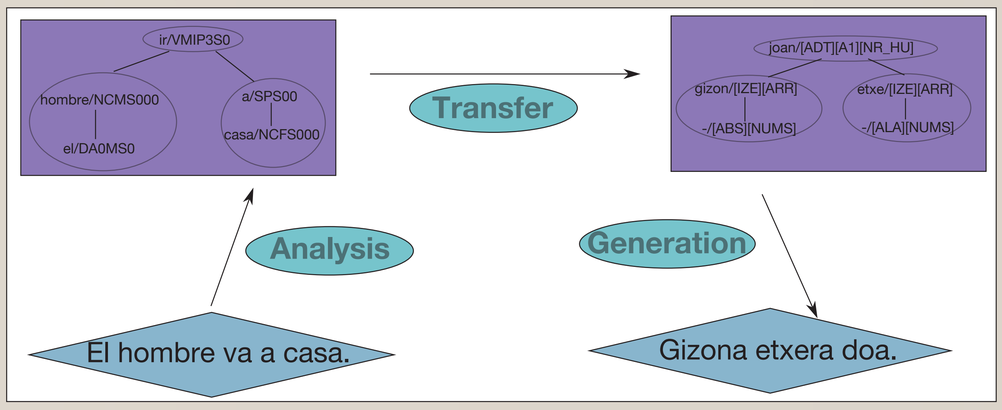
- Matrex system is a **Data-Driven system**.
- Matrex has a **language-independent** design.
- **SMT aligned phrases are enriched** with lingüisticaly motivated aligned chunks.
  Spanish is divided in chunks using *Freeling* shallow parser.
  Basque is divided in chunks using *Eusmg*
  Those chunks are aligned using an "edit-distance style" alignment algorithm.



- **Tunning MaTrEx to Basque**: due to Basque peculiarities slight changes are included.
  1) Basque text is **segmented into morphemes**. for example:
  etxekoan ⟶ etxe<IZE><ARR> +<S><GEL> +<S><INE>
  /in the one of the house/ */house<NOUN>/* */of<SING>/* */in<SING>/*
  2) After translation a **postprocessing stage** is necessary. To carry out the generation we have reused Matxin's lexical generation module.

MATXIN SYSTEM

- Matxin is a **classical transfer system**.
  Divided in three main steps: analysis, transfer and generation.
- It has been specifically designed for Spanish-Basque language pair.



- **Analysis**: the Freeling toolkit is used to carry out the Spanish dependency parsing.
- **Transfer** is divided into lexical-transfer and syntactical-transfer:
  - For lexical transfer dicctionary is used.
  - For syntactic transfer tree tranformation rules has been developed.
- **Generation**, like transfer, is divided in two substeps. Lexical and syntactical generation.
  - In syntactical generation the order of the dependency tree elements is defined.
  - In lexical generation the word forms are generated, adding suffixes with morphological information to the lemmas. Lexicon of the *Eusmg* has been reused.

EVALUATION

CORPORA

- **Training and development corpora**: Both are extracted from Consumer corpus. A collection of 1036 Spanish articles of the Consumer Eroski magazine and their translation into Basque.
- **Test corpora**: Two different corpora are used for evaluation.
  - ConsumerTest, 1500 sentences extracted, as train and development corpora, from Consumer corpus. So it can be considered **"in-domain"**
  - EitbTest corpus, extracted from a different corpus and considered **"out-of-domain"** in evaluation .
  - A subset of 50 sentences from both test corpora are extracted to carry out the human evaluation.

Some statistics from these corpora:

	Sentences	Spanish words	Basque words
Training corpus	51,949	976,720	786,705
Development	1,292	24,755	19,978
ConsumerTest	1,501	34,231	27,278
EitbTest	1,500	36,783	26,857
ConsumerTestHuman	50	746	-
EitbTestHuman	50	692	-

AUTOMATIC EVALUATION

- For automatic evaluation we have calculated the **BLEU** and **NIST** metrics, on both in-domain (ConsumerTest) and out-domain (EitbTest) test corpora.
- Because of the specific nature of Basque, as pointed above, we perform two types of evaluation:
  - **Word-based evaluation (WB)**
  - **Morpheme-based evaluation (MB)**

	ConsumerTest		EitbTest	
	BLEU	NIST	BLEU	NIST
Matxin-WB	6.31	3.66	9.30	3.13
MaTrEx-WB	8.03	3.69	9.02	2.70
Matxin-MB	12.01	4.62	12.76	3.75
MaTrEx-MB	14.48	4.63	6.25	2.89

HUMAN EVALUATION

- For human evaluation we have calculated the **edit-distance metric** (Przybocki et al., 2006) also called **HTER** or Human-targeted Translation Error Rate (Snover et al., 2006), which is defined as the number of modifications a native Basque translator has to make so that the resulting edited translation is an easily understable Basque sentence that constains the complete meaning of the source sentence.
- As in automatic evaluation we have calculated the metrics word based (WB) and morpheme based (MB).

	ConsumerTest	EitbTest
Matxin-WB	43.6	40.4
MaTrEx-WB	57.9	71.8
Matxin-MB	39.1	34.9
MaTrEx-MB	49.6	76.3

CONCLUSIONS

- A **new translation scheme based on morphemes** instead of words.
- **Contradictory results for automatic and human evaluations**, consistent with the findings of (Callison-Burch et al., 2006):
  - The automatic metrics indicate that the data-driven system outperforms the rule-based system on the in-domain data.
    BLEU: difference of 1.68 points (27% relative increase) for word-based and 2.47 points (21% relative increase) for morpheme-based
  - On the contrary, the human evaluation indicates that rule-based system outperforms the data-driven approach for both corpora, irrespective of the corpus.
    HTER "in-domain": 14.3 points for the word-based evaluation and 10.5 points for the morpheme-based evaluation.
    HTER "out-domain": 31.4 points for the word-based evaluation and 41.4 points for the morpheme-based evaluation.
- However, let us to stress that
  - **Matxin has been specifically developed** and designed to translate from Spanish to Basque.
  - **MaTrEx is generic** and the cost of adapting it to Spanish-Basque translation is orders of magnitude lower.
- Next steps:
  - Building a **hybrid system** upon the respective strength of both approaches.
  - Investigating automatic **evaluation metrics** that would be more suited to the evaluation for morpheme-based translation.