MaTrEx: DCU Machine Translation System for IWSLT 2006

Nicolas Stroppa and Andy Way

Dublin City University, School of Computing, NCLT

Outline

DCU@IWSLT 2006

First Participation

- Open Data Track
- Two directions:
  - Italian $\rightarrow$ English
  - Arabic $\rightarrow$ English
- 1-best ASR hypotheses + correct recognition results
- Use the provided training data only

Outline

MaTrEx: A Hybrid EBMT/SMT System

Overview of the system

- A word alignment component (GIZA++)
- A chunking component
- A chunk alignment component
- Two phrase alignment components:
  - "SMT"-style phrase aligner (standard phrase extraction from GIZA++ alignments)
  - "EBMT"-style phrase aligner (phrases are extracted from (i) the chunker and (ii) the chunk aligner)
- A minimum-error rate training component (PHRAMER)
- A decoder (PHARAOH)
- A case and punctuation restoration component

Chunking

Two types of chunking

- Marker-based chunking
  - surface chunking based on marker words
- Treebank-based chunking
  - learner trained on annotated data extracted from treebanks

Marker-Based Chunking

- Approach to EBMT based on the Marker Hypothesis

  "The Marker Hypothesis states that all natural languages have a closed set of specific words or morphemes which appear in a limited set of grammatical contexts and which signal that context." (Green, 1979).

- Universal psycholinguistic constraint: languages are marked for syntactic structure at surface level by closed set of lexemes or morphemes.

Marker-Based Chunking

In my case , it is usally on business , seldom for pleasure .

Marker-Based Chunking

In *my case* , **it** *is usally* **on** *business* , **seldom for** *pleasure* .

- NPs usually start with determiners, or possessive pronouns
- PPs usually start with prepositions

Marker-Based Chunking

In *my case* , **it** *is usally* **on** *business* , **seldom for** *pleasure* .

Nel *mio caso* , **solitamente per** *affari* , **raramente per** *piacere* .

- NPs usually start with determiners, or possessive pronouns
- PPs usually start with prepositions

Marker-Based Chunking

In *my case* , **it** *is usally* **on** *business* , **seldom for** *pleasure* .

Nel *mio caso* , **solitamente per** *affari* , **raramente per** *piacere* .

- NPs usually start with determiners, or possessive pronouns
- PPs usually start with prepositions
- We can use a set of closed-class marker words to segment aligned source and target sentences (determiners, quantifiers, prepositions, conjunctions, possessive pronouns, personal pronouns, punctuation marks)

Treebank-based Chunking

Chunking as a sequence tagging task

Example (using the Inside-Outside-Begin representation)

It	takes	time	to	train	a	train	driver	.
PRP	V	NN	TO	V	DT	NN	NN	.
B-NP	B-VP	B-NP	B-VP	I-VP	B-NP	I-NP	I-NP	O
[ ]$_{NP}$	[ ]$_{VP}$	[ ]$_{NP}$	[	]$_{VP}$	[		]$_{NP}$	

- Tagged data can be extracted from Treebanks (cf. CoNLL 2000 shared task)
- Sequence tagging is performed using a classifier (sliding window)
- Efficient classifiers such as Support Vector Machines (SVM) can be applied (implemented in e.g. YAMCHA)

Chunking

- The lists of English and Italian marker words were extracted from CELEX and MORPHIT respectively, and edited manually.
- We trained YAMCHA on the English and Arabic Penn Tree Banks.

Chunk Alignment

The chunks obtained from the chunkers have to be aligned

English: *[it felt okay] [after the game] [but then] [it started turning black-and-blue] [is it serious ?]*
Italian: *[era a posto] [dopo la partita] [ma poi] [ha cominciato] [a diventare livida] [è grave ?]*

English: *[in my case] [it is usually] [on business] [seldom] [for pleasure]*
Italian: *[nel mio caso] [solitamente] [per affari] [raramente] [per piacere]*

Chunk Alignment Strategies

We assume that for a pair of aligned chunked sentences (e_1^k, f_1^l), we have access to $\mathbb{P}(e_i|f_j)$ and $\mathbb{P}(f_j|e_i)$.

Several alignment strategies

- Edit-distance-like alignment
- Edit-distance-(with jumps)-like alignment
- *IBM model-1-like alignment*

We perform the alignments in both directions and keep the intersection.

Chunk Alignment

How to compute $\mathbb{P}(e_i|f_j)$?

We can use the following information:

- Marker Tags
- Cognate Information
- Word Translation Probabilities (IBM model 1-like)

We can combine the different sources of knowledge within a log-linear model

Remark

All the parameters are computed "on the fly".

Combining the "SMT" and "EBMT" chunks

Hybridity

- EBMT and SMT aligned chunks are merged (counts are added)

Adding EBMT chunks to the SMT chunks:

- adds good alignments which are not present otherwise (less constrained strategy than the phrase extraction heuristic)
- "boosts" already present SMT chunks, i.e. contributes to the direct re-estimation of phrases (rare phrases are over-estimated)

Outline

Data and Preprocessing

- Training was performed using the provided data (no external data)
- We used the OpenNLP tokenizer (a Maximum-Entropy approach) for tokenizing the English and Italian data (a set of regular expressions was added for Italian), and ASVM for Arabic
- Chunking was done with the marker-based chunker (English, Italian), and ASVM (Arabic)
- Chunk Alignment was performed using the Edit-Distance-like aligner (Italian→English), and the Edit-Distance-with-jumps-like aligner (Arabic→English)
- 3-gram Language Model, with Kneser-Ney smoothing
- Minimum-Error Rate Training on dev4 dataset
- Punctuation and case information was restored using SRILM (hidden-ngram and disambig)
- Removed the words in the output that were directly copied from the input

Results - Arabic

ASR (1-best)

	BLEU	*NIST*	*Meteor*	*WER*	*PER*
Official	0.145	4.531	0.402	0.7027	0.5949
Additional	0.1391	4.794	0.4	0.7165	0.5870

Correct Recognition Result

	BLEU	*NIST*	*Meteor*	*WER*	*PER*
Official	0.1624	4.89	0.4336	0.686	0.5678
Additional	0.1589	5.29	0.432	0.6935	0.5537

Results - Italian

ASR (1-best) - Baseline

	BLEU	*NIST*	*Meteor*	*WER*	*PER*
Official	0.2399	6.39	0.5378	0.60	0.49
Additional	0.2568	6.98	0.54	0.59	0.46

Results - Italian

ASR (1-best) - Baseline (B) vs. Matrex (M)

	BLEU	*NIST*	*Meteor*	*WER*	*PER*
Official-B	0.2399	6.39	0.5378	0.60	0.49
Official-M	**0.2598**	**6.59**	**0.5497**	**0.5835**	**0.4869**
Additional-B	0.2568	6.98	0.54	0.59	0.46
Additional-M	**0.2783**	**7.228**	**0.5495**	**0.5662**	**0.4498**

About 2 BLEU points improvement

Results - Italian

Correct Recognition Result - Baseline

	BLEU	*NIST*	*Meteor*	*WER*	*PER*
Official	0.2399	6.39	0.5378	0.60	0.49
Additional	0.2568	6.98	0.54	0.59	0.46

Results - Italian

Correct Recognition Result - Baseline (B) vs. Matrex (M)

	BLEU	*NIST*	*Meteor*	*WER*	*PER*
Official-B	0.2882	7.223	0.6219	0.5528	0.43548
Official-M	**0.3126**	**7.546**	**0.6246**	**0.5315**	**0.4286**
Additional-B	0.3219	8.046	0.6220	0.5188	0.3788
Additional-M	**0.3467**	**8.358**	**0.6245**	**0.4964**	**0.3744**

More than 2 BLEU points improvement

Summary

- We introduced the MaTrEx Data-Driven MT system being developed at Dublin City University
- We presented a method to extract aligned phrases using chunkers and chunk akigners:
  - Marker-based chunking, SVM-based chunking
  - Edit-Distance-like chunk aligners
- We participated in the OpenData Track, for the Italian-to-English and Arabic-to-English directions

Ongoing and Future Work

- Perform a more systematic comparison of the different chunking and alignment strategies
- Insert a segmentation probability directly in the decoding process, in order to give a preference to the phrases that are chunks according to the chunker
- Insert chunk label information in a factored model

Thank you

Thank you for your attention

http://www.computing.dcu.ie/research/nclt