

FACTORED PHRASE-BASED STATISTICAL MACHINE TRANSLATION

Dan Tufiș, Alexandru Ceașu

tufis@racai.ro, aceausu@racai.ro



***Research Institute for Artificial Intelligence
Romanian Academy***

Outline

- Background
 - Data preparation & statistical models construction
 - Experiments with simple statistical translation (several language pairs)
- Factored translation (RO \leftrightarrow EN)
 - Factors of the translation models
 - Evaluation of the factored systems
- Conclusions
- Further work

Background

- A small project SEE-ERA.net (2007-2008)
 - 9 months, 20,000 EURO
- A feasibility study with respect to the MT R&D for some South Slavic and Balkan languages (XX: Bulgarian, Greek, Romanian, Serbian and Slovene)
- Partners: Research Institute for Artificial Intelligence - RACAI (Romania), Institute for Bulgarian Language - IBL (Bulgaria), Institute for Language and Speech Processing - ILSP (Greece), Faculty of Mathematics - MATF (Serbia), Jožef Stefan Institute - JSI (Slovenia)
- Objectives:
 - Collecting parallel texts in the partners languages
 - Preprocessing the parallel texts, validating them and building a pilot parallel training corpus for statistical MT
 - Building language models and translation models (XX<->EN)

What we achieved (1)

- Two multilingual corpora 1-1 sentence aligned to the hub language (English):
 - SEnAC (SeaEra.net Administrative Corpus)
 - languages: BG, EL, RO, SL
 - source: JRC Acquis-Communautaire corpus
 - size: 60,389 sentences in each language (more than 1,2 Mio words/language)
 - for the purpose of harmonizing this work with the objectives of larger projects of some partners we included some non-SSB languages (CZ, FR, DE and EN as the hub language)
 - SEnLC (SeaEra.net Literary Corpus)
 - languages: BG, EL, RO, SR, SL
 - source: Jules Verne's novel “Around the world in 80 days”
 - size: 4409 sentences in each language (about 60,000 words/language)

What we achieved (2)

- Heavy annotation of the two corpora for
 - At least tokenization, tagging, lemmatization
 - Additionally (RO) chunking, dependency linking
- Word and phrase alignment of the SEnAC
 - For RO-EN we used an in-house reified word-aligner (COWAL)
 - For other XX-EN we used GIZA++
- An Alignment Editor allows the visualization and the corrections of the word alignments
- Language models for the concerned SSB languages
- Different Factored Translation Models (both directions) between EN and BG, EL, RO, SL
- Started very encouraging MT experiments using MOSES, proving that developing reasonably accurate MT systems for the concerned languages is realistic and, provided adequate language resources are available, could be done within acceptable time limits.

Example of a TU encoding in SEnAC

<tu id="3936">

<seg lang="bg">

<s id="31985L0337.n.83.1">

Резултатите от консултациите и информацията , събрана съгласно членове 5,6 и 7 , трябва да се вземат предвид при процедурата по издаването на разрешението .</s></seg>

<seg lang="el">

<s id="31985L0337.n.85.1">

Οι πληροφορίες που συγκεντρώνονται δυνάμει των άρθρων 5 , 6 και 7 πρέπει να λαμβάνονται υπόψη στα πλαίσια της διαδικασίας για τη χορήγηση αδείας .</s></seg>

<seg lang="en">

<s id="31985L0337.n.84.1">

Information gathered pursuant to Articles 5 , 6 and 7 must be taken into consideration in the development consent procedure .</s></seg>

<seg lang="ro">

<s id="31985L0337.n.83.1">

Informațiile culese conform art. 5 , 6 și 7 trebuie să fie luate în considerare în cadrul procedurii de autorizare .</s></seg>

<seg lang="sl">

<s id="31985L0337.n.83.1">

Informacije , zbrane skladno s členi 5 , 6 in 7 , se morajo upoštevati v postopku za pridobitev soglasja za izvedbo .</s></seg>

</tu>

```
<seg lang="ro">
```

```
<s id="jrc31970D0244-ro.7.2">
```

```
<w lemma="acest" ana="Dd3fpr---e" chunk="Np#1">Aceste</w>
```

```
<w lemma="previziune" ana="Ncfp-n" chunk="Np#1">previziuni</w>
```

```
<w lemma="fi" ana="Vaip3p" chunk="Vp#1" >sunt</w>
```

```
<w lemma="repartiza" ana="Vmp--pf" chunk="Vp#1">repartizate</w>
```

```
<w lemma="pe" ana="Spsa" chunk="Pp#1">pe</w>
```

```
<w lemma="categorii" ana="Ncfp-n" chunk="Pp#1,Np#2">categorii</w>
```

```
<w lemma="de" ana="Spsa" chunk="Pp#2">de</w>
```

```
<w lemma="cheltuială" ana="Ncfp-n" chunk="Pp#2,Np#3"
```

```
>cheltuieli</w>
```

```
<c>.</c>
```

```
</s>
```

Example of a RO sentence encoding

[Aceste previziuni]_{Np#1} [sunt repartizate]_{Vp#1} [pe [categorii]_{Np2}]_{Pp1} [de [cheltuieli]_{Np3}]_{Pp2}

**Example of a
RO sentence
encoding
and
dependency
linking**

<seg lang="ro">

<s id="jrc31970D0244-ro.7.2">

<w lemma="acest" ana="Dd3fpr---e" chunk="Np#1">Aceste</w>

<w lemma="previziune" ana="Ncfp-n" chunk="Np#1"
head="0">previziuni</w>

<w lemma="fi" ana="Vaip3p" chunk="Vp#1" head="3">sunt</w>

<w lemma="repartiza" ana="Vmp--pf" chunk="Vp#1"
head="1">repartizate</w>

<w lemma="pe" ana="Spsa" chunk="Pp#1" head="5">pe</w>

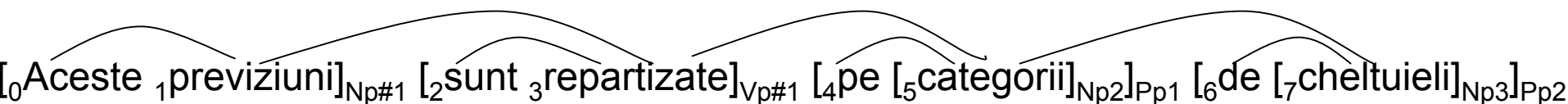
<w lemma="categorii" ana="Ncfp-n" chunk="Pp#1,Np#2"
head="3">categorii</w>

<w lemma="de" ana="Spsa" chunk="Pp#2" head="7">de</w>

<w lemma="cheltuială" ana="Ncfp-n" chunk="Pp#2,Np#3"
head="5">cheltuieli</w>

<c>.</c>

</s>



Word Alignment in Four Language Pairs

The image displays four instances of the 'Align Viewer' application, each showing a word alignment task for a specific language pair. Each window includes a list of source words on the left, a list of target words on the right, and a 'Word Properties' panel on the right side. The 'Translation unit' is set to 69 in all cases.

Window 1: jrcorpus-seeeranet-en-sl-sgml.align

Translation unit: 69

Source Word	Source Index	Target Word	Target Index
The	1	Uradni	1
Official	2	list	2
Journal	3	Skupnosti	3
of	4	se	4
the	5	objavi	5
Community	6	v	6
shall	7	štinh	7
be	8	uradnih	8
published	9	jezikih	9
in	10	.	10
the	11	.	11
four	12	.	12
official	13	.	13
languages	14	.	14
.	15	.	15

Word Properties: Ana, Ncnsan

Property	Value
Chunk	list
Lemma	list
Ortho	list
Position	2

Ready.

Window 2: jrcorpus-seeeranet-en-ro-sgml.align

Translation unit: 69

Source Word	Source Index	Target Word	Target Index
The	1	Jurnalul	1
Official	2	Oficial	2
Journal	3	al	3
of	4	Comunității	4
the	5	apare	5
Community	6	în	6
shall	7	cele	7
be	8	patru	8
published	9	limbi	9
in	10	oficiale	10
the	11	.	11
four	12	.	12
official	13	.	13
languages	14	.	14
.	15	.	15

Word Properties: Ana, Ncnsry

Property	Value
Chunk	Np#1
Lemma	jurnalul
Ortho	Jurnalul
Position	1

Ready.

Window 3: jrcorpus-seeeranet-en-el-sgml.align

Translation unit: 69

Source Word	Source Index	Target Word	Target Index
The	1	Η	1
Official	2	Επίσημη	2
Journal	3	Εφημερίδα	3
of	4	της	4
the	5	Κοινότητας	5
Community	6	εκδίδεται	6
shall	7	στις	7
be	8	τέσσερις	8
published	9	επίσημες	9
in	10	γλώσσες	10
the	11	.	11
four	12	.	12
official	13	.	13
languages	14	.	14
.	15	.	15

Word Properties: Ana, Ncfsn

Property	Value
Chunk	
Lemma	εφημερίδα
Ortho	Εφημερίδα
Position	3

Ready.

Window 4: jrcorpus-seeeranet-en-bg-sgml.align

Translation unit: 69

Source Word	Source Index	Target Word	Target Index
The	1	официалният	1
Official	2	вестник	2
Journal	3	на	3
of	4	общността	4
the	5	се	5
Community	6	издава	6
shall	7	на	7
be	8	четирите	8
published	9	официални	9
in	10	езика	10
the	11	.	11
four	12	.	12
official	13	.	13
languages	14	.	14
.	15	.	15

Word Properties: Ana, Ns0

Property	Value
Chunk	
Lemma	вестник
Ortho	вестник
Position	2

Ready.

The Noisy Channel Model of Translation

$$\text{Target}^* = \underset{\text{Target}}{\operatorname{argmax}} P(\text{Source} | \text{Target}) * P(\text{Target})$$

- The seminal work of the IBM group (the famous 5 models)
- $P(\text{Source}|\text{Target}) \times P(\text{Target})$ allows us to have a sloppy translation model
- Hopefully $P(\text{Target})$ will correct the mistakes of the translation model

The Fundamental Ingredients

1. A Language Model of the target language: $P(\text{Target})$
 - Measures fluency
 - Probability of an Target sentence
 - Provides the right word ordering and collocations
 - Provides a set of fluent sentences to test for potential translation
2. A Translation Model: $P(\text{Source}|\text{Target})$
 - Measures faithfulness
 - Probability of an (Source, Target) pair
 - Provides the right bag of words
 - Tests if a given fluent sentence is a translation
3. A Decoder: argmax
 - An effective and efficient search technique to find Target*
 - Usually this is a heuristic search program

Simple SMT Experiments

- Used the MOSES environment for single factored (wordform) translation
- Language models and translation models built on word surface forms (no available pre-processing tools for BG, EL, SL)
- Language pairs: RO->EN, EN->RO, SL->EN, EN->SL, EN->EL, EL->EN, EN->BG, BG->EN

Evaluation of the See-ERA.net

LANGUAGE PAIR	NIST score	BLEU-1 score
English to Romanian	4.9923	0.3634
Romanian to English	4.6191	0.3862
English to Greek	3.9730	0.3005
Greek to English	3.8036	0.3533
English to Slovene	3.6719	0.2293
Slovene to English	4.0589	0.2450

Factored Translation Models for RO->EN & EN->RO (STAR project)

- Combination of various factors of translation elements:
 - Wordforms
 - Lemmas
 - Morpho-Syntactic Descriptors (MSDs)
 - Corpus Tags (POS)
 - Chunking and linkage (not yet, to be added as future factors for FTMs)
- JRC-Acquis (DGT-TM) plus RoWordNet & EnWordNet
- Phrase-Based Statistical Translation
- Used the full power of the MOSES environment (Philipp Koehn)
- Estimation of LMs and TMs judged in terms of perplexity
- MOSES Decoder allows for multiple TMs (linear interpolation of multiple factors- $h(t,s)$; the λ s are computed by Och's MERT algorithm)

$$t^* = \arg \max_e \sum_{k=1}^k \lambda_k h_k(t, s)$$

Factoring the translation

1. Translating lemmas (lemma/lemma factor of the TM);
2. Generating the morpho-syntactic descriptions candidates for each translated lemma (lemma/MSD ambiguity class from the target LM);
3. Translating source POSs into target POSes (POS/POS factor of the TM);
 - 3.1. Mapping POSes into MSDs (POS/MSD recovery in the target language);
 - 3.2. Computing the most likely MSDs for the translated lemmas by intersection of the candidates from step 2 and the result of step 3.1
4. Generating the target surface form candidates based on translated lemma of step 1 and morpho-syntactical description of step 3.2 (lemma/MSD associations from the target LM)
5. Reordering the words in the target language combining two information sources from the target LM: wordform 3grams & MSD 3grams (in the future, adding linking and chunking)

Evaluation of the STAR

Done in the standard way:

- Training data (925,032 translation units, 32,730,079 tokens)
- Tuning data (MERT tuning on 200 translation units)
- Evaluation on data (1,000 sentences) not seen during the training or tuning using the official NIST tool available at:
<ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-v12.pl>

The most recent evaluation results shown below, obtained by Al. Ceașu in his PhD thesis (successfully defended on 18th May)

LANGUAGE PAIR	NIST score	BLEU-1 score
English to Romanian	9.6634	0.4722
Romanian to English	11.5691	0.5276

Research Institute for Artificial Intelligence

Xml Web Services

Factored Translation

Language pair:

- English-Romanian
 Romanian-English

Input text:

```
This Regulation shall apply to shipments of waste:
(a) between Member States, within the Community or with transit through third countries;
(b) imported into the Community from third countries;
(c) exported from the Community to third countries;
(d) in transit through the Community, on the way from and to third countries.
```

Show intermediate results

Translate

Annotated input text (TextProcessingWebService):

```
This|this|DMS|Dd3-s Regulation|regulation|NN|Ncns shall|shall|AUX|Vaip apply|apply|VINF|Vmn
to|to|PREP|Sp shipments|shipment|NNS|Ncnp of|of|PREP|Sp waste|waste|NN|Ncns :|:|COLON|COLON
(|(|LPAR|LPAR a|a|TS|Ti-s )|)|RPAR|RPAR between|between|PREP|Sp Member|member|NN|Ncns
States|state|NNS|Ncnp ,|,|COMM&|COMMA within|within|PREP|Sp the|the|DM|Dd Community|community|NN|Ncns
or|or|CCOO|Cc-n with|with|PREP|Sp transit|transit|NN|Ncns through|through|PREP|Sp
third_countries|third_countries|MNP|Np ;|;|SCOLON|SCOLON
```

Output:

```
prezentul regulament se aplică transporturilor de deșeuri :
( a ) între statele membre , în cadrul comunității sau cu tranzitarea altor țări terțe ;
( b ) importate în comunitate din țări terțe ;
( c ) exportate din comunitate către țări terțe ;
( d ) în tranzit prin comunitate , pe drumul de_la și către țări terțe .
```

Web Services

- Home
- Text Processing
- Factored Translation
- Language Identification

Factored Translation

- Web Service Description



Research Institute for Artificial Intelligence Xml Web Services

Factored Translation

Language pair:

- English-Romanian
 Romanian-English

Input text:

Prezentul regulament stabilește măsurile ce trebuie adoptate pentru monitorizarea comerțului între Comunitate și țări terțe cu substanțe frecvent folosite la fabricarea ilicită a stupefiantelor și substanțelor psihotrope , cu scopul de a împiedica deturnarea acestor substanțe .

Show intermediate results

Translate

Annotated input text (TextProcessingWebService):

Prezentul|prezent|NSRY|Ncmsry regulament|regulament|NSN|Ncms-n stabilește|stabili|V3|Vmip3s
 măsurile|măsură|NPRY|Ncfpry ce|ce|RELR|Pw3--r trebuie|trebui|V3|Vmip3s adoptate|adopta|VPPF|Vmp--pf
 pentru|pentru|S|Spsa monitorizarea|monitorizare|NSRY|Ncfsry comerțului|comerț|NSOY|Ncmsoy între|
 între|S|Spsa Comunitate|comunitate|NSRN|Ncfsrn și|și|CR|Crssp țări|țară|NPN|Ncfp-n
 terțe|terț|APN|Afpfp-n cu|cu|S|Spsa substanțe|substanță|NPN|Ncfp-n frecvent|frecvent|R|Rgp
 folosite|folosit|APN|Afpfp-n la|la|S|Spsa fabricarea|fabricare|NSRY|Ncfsry

Output:

this regulation establishes measures to be taken to monitor trade between the community and third_countries by substances frequently used in the illicit manufacture of narcotic drugs and psychotropic_substances , with the aim of preventing the diversion of these substances .

Web Services

- Home
- Text Processing
- Factored Translation
- Language Identification

Factored Translation

- Web Service Description

Conclusions (1)

Conducted experiments

- **RO-EN & EN-RO** (multi-factored translation)
- **SL-EN & EN-SL** (mono-factored translation)
- **EL-EN & EN-EL** (mono-factored translation)

On-going experiments

- **BG-EN** (mono-factored translation)
- **RO-DE** (mono-factored translation)
- **RO-FR & FR-RO** (multi-factored translation)

The use of linguistic knowledge, significantly improve the translation quality!

Conclusions (2)

- With the current technologies building a domain specific SMT is fast;
- Accuracy (at least for assimilation purposes) is better and better, strongly competing and even exceeding the previous rule-based systems;
- As more and more parallel corpora are available, more and more (small) research groups come closer to the performances of Google, Yahoo, Microsoft and other indexers of the web;
- In order to get over the performances of the owners of zillions of parallel raw data, more linguistic knowledge should be used. The required volume of linguistically preprocessed data (for the same performance) is smaller by several orders of magnitude; yet, preprocessing is not easily available for all languages!
- BLARK concept (promoted by the CLARIN ERI) should be an open door towards this aim;
- A new generation of “decoders” will probably make the next boost. Our proposal is for a reified translation modeling;

Further Work (1)

Factorisation versus Reification

- In the MOSES implementation of the factored translation process, each factor adds translation hypotheses to the search space, independently of the other factors. The decoder has the role to make the best choice among the generated translation hypotheses;
- In a reified translation model, a translation is built iteratively, segment by segment; the choice of a particular segment is selected considering *simultaneously* all the factors relevant for the respective segment, thus making the decoding phase unnecessary. The translated phrases are added in several iterations, beginning with the highest confidence fragments and using them, in the subsequent steps, as constraints for generating new fragments. Reification could include dependency linking much easier than factorization;

Further Work (2)

Going beyond the sentence level

– Translation for several language pairs requires access to the contextual information:

- “ε Mergea în zig-zag ...”;



“Era albă”->”? (It, She) was white” ->”Era alb(ă)?”

? (He? She? It?)

Pro-drop languages, anaphoras

– Formal/colloquial language:

you->? (tu, dumneata, dumneavoastră, voi)

- Honorifics (“You” referring to your boss normally should be translated with “Dumneavoastră”; “You” referring to your relative should be translated with “Tu” or maybe with “Dumneata”)

– Tense/mood/aspect (TMA problem)

Further Work (3)

- **Exploiting comparable corpora for enhancing the translation models**
- **Investigating new evaluation methods for the translation quality allowing to combine fragments from multiple translation engines output**
- **Combining statistical and rule-based strategies for translation generation**
- **Exploring new triangulation methods (translation using a pivot language S->P->T)**

These topics are addressed in two FP7 ICT-call4 proposals!

Thank you!

Q/A ?

