Integrating Multi Word Expressions in Statistical Machine Translation

- From **Zied Eloumi**'s Master Degree Presentation
- Supervised by Olivier Kraif and Laurent Besacier

MWEs and MT

MWEs are a challenge for MT

- Example1 : **idioms**

« Ils étaient ***tombés dans les pommes*** sur un lit. »

Google-TR :« They had ***fallen in apples*** on a bed. »

-> s'évanouir …

- Example 2 : **collocations**

« L’honorable parlementaire me ***fait*** trop d’***honneur***.»

Google-TR « The honorable member ***made*** me too much ***honor***.»

MWEs and MT

- Example3 : ***phrasal verbs***

« Surely they must ***call*** the operation ***off*** now? »

Google-TR :« Certes, ils doivent ***appeler*** l'opération maintenant? »

Master Contributions

- Extracting MWEs (auto. or semi auto.)
- Propose specific corpora for evaluating MT of MWEs
- Propose efficient approaches to handle MWEs in MT

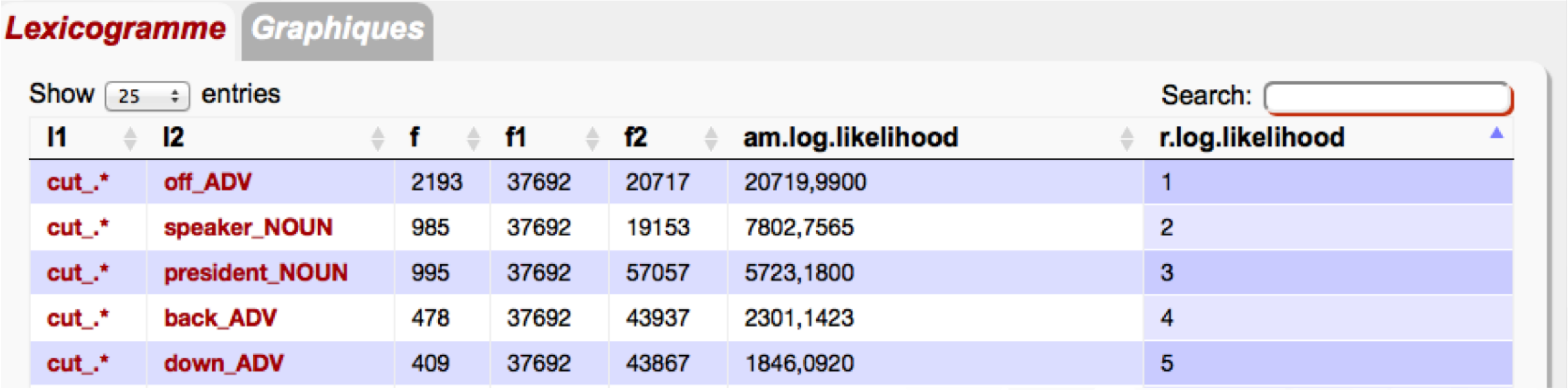
1. Extraction of a specific corpus for evaluating MT of MWEs

Semi Automatic Tool for Extraction of phrasal verbs

Using LIDILEM's tool - Lexico-gramme [*O. Kraif et S. Diwersy (2012)]*

Exemple : Pivot=*cut*

Europarl, News and Emolex (litterary texts) corpora are indexed



Show 25 entries

Search:

l1	l2	f	f1	f2	am.log.likelihood	r.log.likelihood
cut_.*	off_ADV	2193	37692	20717	20719,9900	1
cut_.*	speaker_NOUN	985	37692	19153	7802,7565	2
cut_.*	president_NOUN	995	37692	57057	5723,1800	3
cut_.*	back_ADV	478	37692	43937	2301,1423	4
cut_.*	down_ADV	409	37692	43867	1846,0920	5

Extraction of phrasal verbs (lexico-gramme)

MWEs Selection

- **FR-EN:** Collocations with word distance from 0 to 3

  show stats in separate file

- **EN-FR:** “Phrasal Verbs” with word distance from 0 à 5

  show stats in separate file

- Keep subset corpus with high density of MWEs

En-Fr Corpus (1)

- 500 sentence pairs
- 40 different phrasal verbs .
- 74 different idioms

	#Corpus EN-FR	#Phrasal verbs	#Idioms
Litterary Texts	272	189	83
Europarl	213	163	50
News	15	14	1
TOTAL	**500**	**366**	**134**

En-Fr Corpus (2)

- Sample sentences
- Hell, man, I'd have called it off if you hadn't shown

  Je te jure, j'aurais tout annulé si tu n'étais pas venu
- He picked up a remote and turned the stereo on

  S'emparant d'une télécommande, il alluma la stéréo
- He says you have important work to do and no more time to fool around

  Il dit que tu as un travail important à accomplir et plus de temps à perdre
- Clint was dead tired and about to fall asleep on his feet

  Clint, épuisé, aurait pu s'endormir debout
- Some guys putting the scares on

  Des mecs venus jouer les gros bras

Fr-En Corpus (1)

- 500 sentence pairs
- 40 collocations : NOUN+VERB , VERB+NOUN , Phrasal verbs
- 20 idioms

	#Corpus FR-EN	#Phrasal verbs	#Idioms
Litterary Texts	253	201	52
Europarl	235	191	44
News	12	8	4
TOTAL	**500**	**400**	**100**

Fr-En Corpus (2)

- Sample sentences
- L' honorable parlementaire me fait trop d' honneur.

  The honourable Member is being too kind to me.
- Faites cela, et ne m'adressez plus aucune parole.

  Do this and speak no more to me.
- Je souhaiterais tirer quelques conclusions supplémentaires.

  I should like to formulate a few more conclusions.
- Aucun des deux ne lui tendit la perche ; ils attendirent patiemment.

  Neither offered to help him , but both waited patiently.
- Le garçon accusa le coup.

  The boy looked shocked.

2. Handling MWEs in MT

LIG (moses) baseline *vs* Google-TR

	Corpus of 500 sent.		Detail per corpus or per category on Tst-MWE				Idioms only in Tst-MWE #134 sent
	Tst-NORMAL	Tst-MWE	PV only #366 sent	Litt. Texts #272 sent	Europarl #213 sent	News #15 sent	
Moses-LIG	24.87%	20.83%	22.72%	12.59%	29.33%	21.94%	15.21%
Google-TR	19,27%	19,81 %	18.67%	10,9%	21,82%	12,00%	19.75%

BLEU scores (Moses-LIG vs Google)

“Witness” corpus
With 500 randomly
selected sentences

Our corpus with higher density of MWEs (500 sentences)

Pre-processing (of PV) for En-Fr SMT training

- Gluing verb and prep for training / test
- Hell, man, I'd have called it off if you hadn't shown

  => Hell, man, I'd have **called_off** it if you hadn't shown
- He picked up a remote and turned the stereo on

  => He picked up a remote and **turned_on** the stereo
- Problems

  =>This pre-processing is not straightforward (need for MWE detection)

Pre-processing (of PV) : Results

	Corpus of 500 sent.		Detail per corpus or per category on Tst-MWE				Idioms only in Tst-MWE #134 sent
	Tst-NORMAL	Tst-MWE	PV only #366 sent	Litt. Texts #272 sent	Europarl #213 sent	News #15 sent	
Moses-LIG	24.87%	20.83%	22.72%	12.59%	29.33%	21.94%	15.21%
+pre-proc	23.81 %	21.19%	23.18%	14.10%	29.38%	19.80%	14.79%

BLEU scores (Moses-LIG vs Google)

Pre-processing (of PV) : Results

	Corpus of 500 sent.		Detail per corpus or per category on Tst-MWE				Idioms only in Tst-MWE #134 sent
	Tst-NORMAL	Tst-MWE	PV only #366 sent	Litt. Texts #272 sent	Europarl #213 sent	News #15 sent	
Moses-LIG	24.87%	20.83%	22.72%	12.59%	29.33%	21.94%	15.21%
+pre-proc	23.81 %	21.19%	23.18%	14.10%	29.38%	19.80%	14.79%

BLEU scores (Moses-LIG vs Google)

- Slight but not significant improvement
- BLEU is calculated on the whole corpus => do not see the effect on smaller events
- Need an evaluation metric that could “zoom” on MWEs

Processing of idioms

- Using a pre-defined « dictionary » of idioms
- Constrained SMT decoding
- Example :

  And you thought my job was a piece of cake !

  And you thought my job was <idiom translation="***facile***">***a piece of cake***</idiom>

  And you thought my job was <idiom translation="***fastoche***">***a piece of cake***</idiom>

  And you thought my job was <idiom translation="***une sinécure***">***a piece of cake***</idiom>
- BLEU goes **from 15.21 % to 30.71 %** on the 134 sentences containing idioms but...
- We used a small pre-defined list of idioms that covered 100 % of our test set => not very realistic
- Idiom detection is not addressed here (pre-defined list)

Conclusion

- Corpora with bigger density of MWEs than in usual text
- Need for an automatic metric that focuses on specific events such as MWEs (otherwise we do not see much differences)
- First propositions to handle MWE
  - -Pre-processing of PV for En-Fr SMT
  - -Use constrained decoding for Idioms
- Example

Source	surely they must *call* the operation *off* now ?
Reference	maintenant, ils doivent sûrement *annuler* l' opération .
Hyp (baseline)	ils doivent *appeler* l' opération maintenant ?
Hyp (+pre-proc)	ils doivent *annuler* le fonctionnement maintenant ?