How hard is it to automatically translate phrasal verbs from English to French?

Carlos Ramisch, Laurent Besacier, Alexander Kobzar

CAMELEON workshop - 13/12/2013
(originally presented at MUMTTT 2013)

Outline

Outline

1. Why are phrasal verbs hard to translate?
   - Definitions
   - Characteristics
   - Scope and goals
2. How to measure translation quality?
   - Evaluation protocol
   - Annotation
3. How hard is it to automatically translate phrasal verbs?
   - Experimental setup
   - Results
4. Wrapping up

Phrasal verbs (PVs)

Main verb + prepositional or adverbial particle

take + *on* or *away* =
take on $\Longrightarrow$ *take on a challenge*
take away $\Longrightarrow$ *I take away your books*

Examples

take off, *give up*, *pull out*, *pick up*, *make up*, *sort out*

Why phrasal verbs?

- Frequent type of MWE in (spoken) English [Sinclair, 1989]
- Complex syntactic and semantic behaviour
- Language asymmetries (*take off* $\Longrightarrow$ *décoller*)

Syntactic characterisation

Intransitive

she will show up later, the aircraft takes off

Transitive

Joint: *we made up this story, he took off his shoes*
Split: *we made this story up, he took them off*

- ***Verb-particle constructions***: *put off, give up, move on*
- ***Prepositional verbs***: *talk about, rely on, wait for*

Semantic characterisation [Bolinger, 1971]

Compositionality range

- ***Literal or compositional***: *take away*
- ***Aspectual or semi-idiomatic***: *fix up*
- ***Idiomatic***: *pull off*

Senses

- *figure out*, *look up*: 1 WN sense
- *pick up*: 16 WN senses
- *break up*: 19 WN senses

Our focus

Split transitive phrasal verbs

Verb + NP + Particle

Examples

take *your clothes* **off**, **give** *it* **up**, **pull** *things* **out**

- Word-based SMT [Brown et al., 1993] → phrase-based SMT [Koehn et al., 2003] → hierarchical SMT [Chiang, 2007]
- SMT can deal with some types of (contiguous) MWEs
- Intuitively, split PVs are hard to model

Our goals

Investigate the following questions:

- What proportion of split phrasal verbs is translated correctly/acceptably by current SMT paradigms?
- Which SMT paradigm, phrase-based or hierarchical, can better handle these constructions?
- What are the main factors that influence translation quality?

Our hypotheses

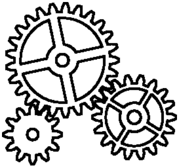
H1 Current SMT systems cannot translate PVs correctly

H2 Hierarchical models provide better translations than phrase-based models

H3 The quality of PV translation is determined mainly by their frequency in training data

Our experiments

1. Train two standard English-French SMT systems
   - phrase-based system (PBS)
   - hierarchical system (HS)
2. Translate a set of English sentences containing split PVs
3. Manually annotate PV translation quality



Outline

Translation quality

How to measure translation quality of a specific linguistic phenomenon inside the sentence?

- BLEU, NIST, Meteor - no insight into errors
- Manual annotation
    1. Specific test set
    2. Carefully designed annotation guidelines
    3. Interface to highlight the target context

1) Test set selection

1. ***Automatic extraction***: Parsing with RASP + mwetoolkit
   - VERB: POS starts with VV
   - OBJECT: sequence of 1 to 5 words, excluding verbs
   - PARTICLE: POS is II, RR or RP and depends on the verb with NCMOD_PART relation
2. ***Filtering heuristics***
   - Particles *about*, *well*, *at*
   - Verbs *go*, *walk*, *do*, *see* + locative words
   - Locative words + *here*, *there*, *way*
   - Verbs with double particles
   - Expressions *upside down*, *inside out*, *all over*
   - Long sentences (> 50 words)
3. ***Manual validation***

2) Annotation guidelines I

- Source and target (no reference) containing 1 PV
- Annotation data:
  - 250 similar PV translations - once
  - 500 different PV translations - consecutively in random order
- Highlight PV and its translation
- `http://cameleon.imag.fr/xwiki/bin/view/Main/Phrasal_verbs_annotation`

2) Annotation guidelines II

Adequacy

Extent to which the meaning of the English PV is preserved in the French translation

- 3 - FULL
- 2 - PARTIAL
- 1 - NONE
- 0 - UNABLE TO JUDGE

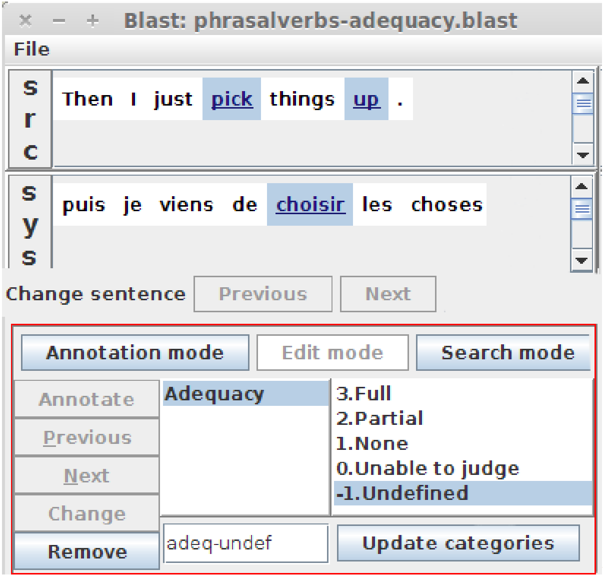
2) Annotation guidelines III

Fluency

Grammatical correctness in French of the translated PV (regardless of meaning)

- 4 - FLUENT
- 3 - NON-NATIVE
- 2 - DISFLUENT
- 1 - INCOMPREHENSIBLE

3) Annotation interface - BLAST [Stymne, 2011]



Annotation process

- 750 sentences, 4 annotators
- Proficient in English and French
- Access to Wordnet and online dictionaries
- Fluency and adequacy annotated in two passes

Inter-annotator agreement

- Pilot dataset of 156 sentences, 5 annotators
- ***Fluency***: $\kappa = 0.50$
- ***Fluency pairwise***: $\kappa = .33$ to $\kappa = .72$
- ***Adequacy***: $\kappa = .35$
- ***Adequacy pairwise***: $\kappa = .23$ to $\kappa = .52$

Outline

1. Why are phrasal verbs hard to translate?
   - Definitions
   - Characteristics
   - Scope and goals
2. How to measure translation quality?
   - Evaluation protocol
   - Annotation
3. How hard is it to automatically translate phrasal verbs?
   - Experimental setup
   - Results
4. Wrapping up

Training and test data

English-French *TED Talks* parallel corpus [Cettolo et al., 2012][1]

	# sentences	
	Sys. 1	**Sys. 2**
Shared training set	137,319	137,319
PVs training set	1,034	1,037
Shared dev. set	2,000	2,000
PVs test set	1,037	1,034
Total	141,390	141,390

[1] https://wit3.fbk.eu/

Machine translation systems

Phrase-based system (PBS) and ***Hierarchical system (HS)***

- Moses toolkit [Koehn et al., 2007], standard training parameters[2]
- Word-alignment: GIZA++
- Phrase table extraction: *grow-diag-final* heuristic
- Language model: 5-grams, IRSTLM, monolingual part of training corpus
- Tuning: MERT
- Print out alignment information

[2] `http://www.statmt.org/moses/?n=Moses.Baseline`

Sanity check

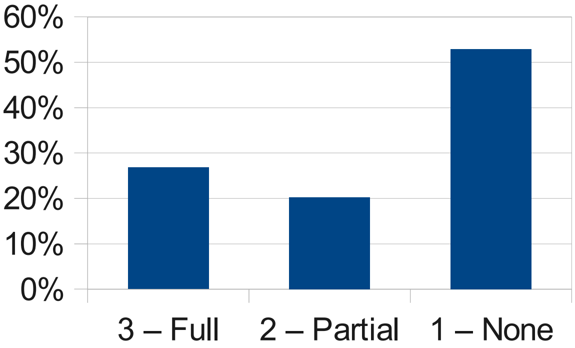
Compare with Google Translate - automatic metrics

BLEU			TER		
PBS	HS	Google	PBS	HS	Google
29.5	25.1	32.3	52.6	55.7	50.3

Translation examples

	could **boil** *this poem* **down** *to saying*
PBS	*pourriez* **furoncle** *ce poème* **jusqu'** *à dire*
HS	*pourriez* **bouillir** *ce poème* **descendu** *à dire*
	he would **think** *it* **through** *and say*
Both	*il* **pense que** *ça* **à travers** *et dire*
	you couldn 't **figure** *it* **out**
HS	*vous ne pouvais pas le* **comprendre**
PBS	*vous ne pouviez pas le* **découvrir**
	Then we 'll **test** *some other ideas* **out**
Both	*puis nous allons* **tester** *certains autres idées*

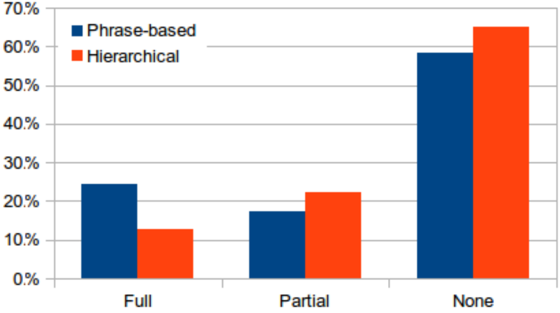
Finding 1: both systems are quite bad



Distribution of annotations - adequacy

Finding 2: PBS is better than HS

	Overall	PBS	HS	PBS-Diff.	HS-Diff.
Fluency (1-4)	2.57	2.67	2.46	2.63	2.25
Adequacy (1-3)	1.73	1.75	1.72	1.65	1.48



Why does the HS perform poorly?

- Some hypotheses
  - Insufficient training material
  - Standard parameters are inadequate
  - In the PBS, generally 2 bi-phrases are enough (*make up*, *make it up*)

Finding 3: similar translations are better

	Similar	Different
Fluency (1-4)	2.82	2.44
Adequacy (1-3)	2.07	1.56

Finding 4: frequency is not the only explanation

Kendall τ rank correlation between PV's average translation quality and:

	Adequacy		Fluency	
	PBS	**HS**	**PBS**	**HS**
Frequency split	0.173	***0.203***	0.168	***0.260***
Frequency joint	0.125	0.096	***0.197***	0.166
Verb-particle distance	0.073	0.061	0.083	0.151

Outline

Wrapping up

Contributions

- A reusable pipeline to build test set (RASP + mwetk + filter)
- An evaluation protocol including detailed guidelines
- A systematic evaluation of EN-FR split PV translation

Wrapping up

Contributions

- A reusable pipeline to build test set (RASP + mwetk + filter)
- An evaluation protocol including detailed guidelines
- A systematic evaluation of EN-FR split PV translation

Main contribution

Even though SMT is a mature framework, flexible constructions like PVs cannot be modelled appropriately

Moving on

- Improve base SMT systems (larger corpora)
- Develop better automatic PV identification
- Conceive clearer guidelines, annotator training, ranking task
- Investigate why the HS performs poorly
- Investigate why PV translation is poor
- Extend to other types of PVs (joint instances, double particles)
- Extend to other types of MWEs (idioms, support-verb constructions)

Thank you! Questions?

Carlos Ramisch, Laurent Besacier, Alexander Kobzar

How hard is it to automatically translate
phrasal verbs from English to French?

Datasets available at
`http://cameleon.imag.fr/xwiki/bin/view/Main/Phrasal_verbs_annotation`

`carlosramisch@gmail.com`

Références I

Bolinger, D. (1971).
The phrasal verb in English.
Harvard UP, Harvard, USA.
187 p.

Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., and Mercer, R. L. (1993).
The mathematics of statistical machine translation: parameter estimation.
Comp. Ling., 19(2):263–311.

Cettolo, M., Girardi, C., and Federico, M. (2012).
Wit^3: Web inventory of transcribed and translated talks.
In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, pages 261–268, Trento, Italy.

Chiang, D. (2007).
Hierarchical phrase-based translation.
Comp. Ling., 33(2):201–228.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007).
Moses: open source toolkit for statistical machine translation.
In *Proc. of the 45th ACL (ACL 2007)*, pages 177–180, Prague, Czech Republic. ACL.

Koehn, P., Och, F. J., and Marcu, D. (2003).
Statistical phrase-based translation.
In *Proc. of the 2003 Conf. of the NAACL on HLT (NAACL 2003)*, pages 48–54, Edmonton, Canada. ACL.

Références II

Sinclair, J., editor (1989).
Collins COBUILD Dictionary of Phrasal Verbs.
Collins COBUILD, London, UK.
512 p.

Stymne, S. (2011).
Blast: A tool for error analysis of machine translation output.
In *Proc. of the ACL 2011 System Demonstrations*, pages 56–61, Portland, OR, USA. ACL.