Improving English-Bulgarian Statistical Machine Translation by Phrasal Verb Treatment

Iliana Simova[1] and Valia Kordoni [2]

[1]Dept. of Computational Linguistics, Saarland University, Germany
ilianas@coli.uni-saarland.de

[2]Dept. of English and American Studies, Humboldt-Universität zu Berlin, Germany
kordonie@anglistik.hu-berlin.de

Sept 3, 2013

Outline

Phrasal Verbs as Multiword Units

1. Verb-particle constructions (VPC)
   verb +
   - intransitive preposition (take *off*)
   - adjective (cut *short*)
   - verb (let *go*)
2. Prepositional Verbs
   verb +
   - transitive preposition (look *for*, refer *to*)

Syntactic Properties of Phrasal Verbs

Syntactic Properties of Phrasal Verbs

- transitivity
  - VPCs - intransitive (*come back*), transitive (*look up* [^{obj} a word])
  - Prepositional Verbs - transitive (*look* [*for* a solution])

Syntactic Properties of Phrasal Verbs

- transitivity
  - VPCs - intransitive (*come back*), transitive (*look up* [^{obj} a word])
  - Prepositional Verbs - transitive (*look* [*for* a solution])
- separability
  - separable VPCs
    (a) She *turned* the light *on*. She *turned on* the light.
    (b) She *turned* it *on*. *She *turned on* it.
  - Inseparable VPCs
    (c) She *fell off* a tree. *She *fell* a tree *off*.
    (d) She *fell off* it. *She *fell* it *off*.

Semantics of Phrasal Verbs

- idiomatic PVs - *do in*, *rain off*
- semi-compositional PVs - *eat up*, *dream on*
- compositional PVs - *take away*, *carry in*

Translation of Phrasal Verbs

Many-to-one mapping (usually):

(1) to put off the decision
da otlozhi reshenieto
to postpone the-decision

(2) to take over peacekeeping operations
da poemat miroopazvashtite operacii
to take-over the-peacekeeping operations

Translation of Phrasal Verbs

Many-to-many mapping:

- 'da'-constructions

  (3) should break off negotiations
  trjabva da prekysne pregovorite
  should (to) interrupt the-negotiations

- reflexive verb particles 'si' and 'se'

  (4) has no intention to give up its plans
  njama namerenie da se otkazva ot planovete si
  has-not intention to give-up-refl from the-plans its

Subtasks

- Data preprocessing
- Creation of a phrasal verb lexicon
- PV identification
- PV knowledge integration
- Evaluation

Language Resources

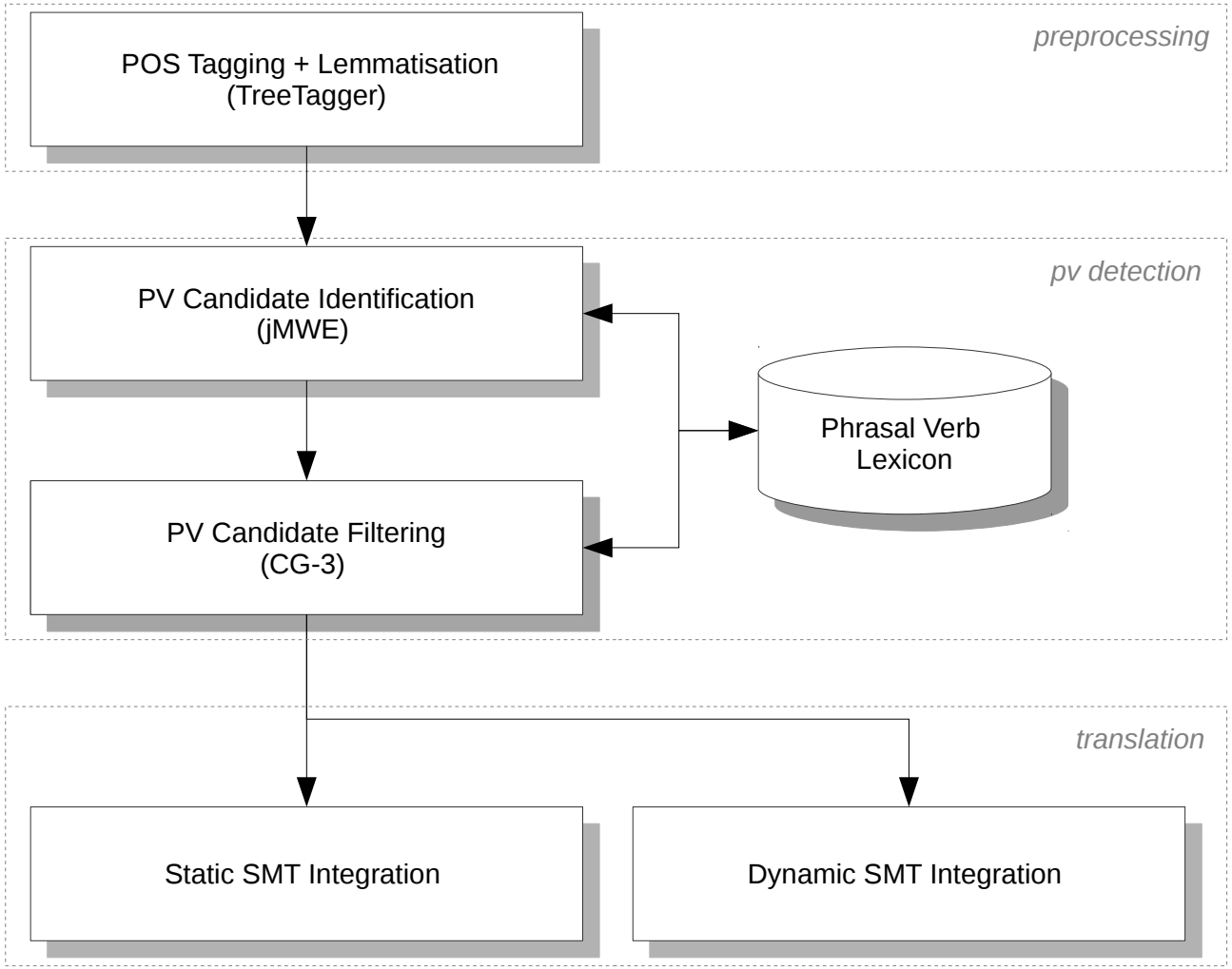
Parallel corpora:

- SETIMES Parallel News Corpus (cleaned version)
  - 151,718 sentence pairs
  - semi-automatically corrected sentence alignments
- Two additional parallel data sets from the EuroMatrixPlus Project
  - 2848 sentence pairs, aligned on the word level by a super-annotator

Monolingual resources:

- Bulgarian National Reference Corpus (a subset of 50 million words)

Pipeline


preprocessing
POS Tagging + Lemmatisation
(TreeTagger)
pv detection
PV Candidate Identification
(jMWE)
Phrasal Verb
Lexicon
PV Candidate Filtering
(CG-3)
translation
Static SMT Integration
Dynamic SMT Integration

Phrasal Verb Lexicon

- Resources
  - → Wiktionary (Category:English_phrasal_verbs)
  - → Phrasal Verb Deamon
  - → WordNet
  - → CELEX
  - → COMLEX
  - → Baldwin VPC dataset
  - → McCarthy dataset
- Additional information
  - → transitivity
  - → separability

Constraint Grammar Formalism

- Ordered set of rules with context-dependent application

  `OPERATION (target) IF (CONTEXT-1)...(CONTEXT-n)`

PV Candidate Filtering with CG-3

- *hold on*, intransitive
  safe: His attempts to hold on until the end of his mandate failed.
  unsafe: He won the elections held on 22 December.
- *take to*, transitive, separable
  safe: Peaceful demonstrators took to the streets this Saturday.
  unsafe: The time it took to establish the full peacekeeping presence.
- *take in*, transitive, separable
  safe: The police took the suspect in for questioning on Friday.
  unsafe: Several weeks have passed since the law took effect in the country.

Translation Experiments

Data sets:

- dev sets (100 sentences)
- tune sets (2000 sentences)
- test sets (800 sentences, 400 with PVs, 400 without)
- train sets (the remaining sentences)

Translation Experiments

Data sets:

- dev sets (100 sentences)
- tune sets (2000 sentences)
- test sets (800 sentences, 400 with PVs, 400 without)
- train sets (the remaining sentences)

Baseline System:

- 5-gram Language Model with SRILM (Bulgarian Referent Corpus)
- Factored Translation Model
  - translation steps:
    EN lemma → BG lemma
    EN POS → BG POS
  - generation steps:
    BG lemma, BG POS → BG word

PV Integration Strategies

- *static*[1] integration
  eat the chocolate up → eat_up the chocolate
- *dynamic*[1] integration

[1] terminology adopted from:

Carpuat, Marine and Mona Diab. 2010. Task-based Evaluation of Multiword Expressions: a Pilot Study in Statistical Machine Translation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 242–245.

Evaluation Results

- PV identification evaluation
- Translation quality evaluation
  - automatic
  - manual

PV Identification Evaluation

Manual evaluations on the test set (800 sentences):

- jMWE+CG-3: Precision = 0.91; Recall = 0.93; F1 = 0.92;
- jMWE: Precision = 0.6;
- PV occurrence statistics

frequency threshold	PV occurrences
≥ 1	107
≥ 5	18
≥ 10	9
≥ 20	4
≥ 30	0

Automatic Evaluation of Translation Quality

Automatic evaluation on (1) 400 sentences containing at least one detected PV, (2) 400 sentences containing no detected PVs, and (3) all 800 sentences.

method	with PVs (1)		no PVs (2)		all (3)	
	bleu	nist	bleu	nist	bleu	nist
baseline	0.244	5.970	0.228	5.726	0.237	6.135
static	0.246	6.016	0.230	5.757	0.239	6.179
dynamic	0.250	5.923	0.226	5.544	0.244	6.020

Examples

original (1); baseline translation (2); static, dynamic strategies (3);

(1) will take over the work of the nato force

(2) shte vzeme vurhu rabotata na sili na nato

will take onto the-work of forces of nato

(3) shte poeme rabotata na silite na nato

will take-over the-work of the-forces of nato

Examples (2)

original (1); baseline translation (2); static (3), dynamic (4);

(1) to leave behind disagreements over iraq

(2) da ostavi zad raznoglasija otnosno irak

to leave in-the-back disagreements over iraq

(3) da preodolejat raznoglasijata nad irak

to overcome the-disagreements over iraq

(4) da ostavi zad gyrba si razlichija po vyprosa za irak

to leave-behind differences on the issue of iraq

Examples (3)

original (1); baseline translation (2); static (3); dynamic (4);

(1) turn themselves in to the international criminal tribunal

(2) se oburna v na mejdunarodnija trubunal

turned-refl inside to the-international tribunal

(3) predade se na mejdunarodnija tribunal

turned-in-refl to the-international tribunal

(4) se oburna v na mejdunarodnija nakazatelen trubunal

turned-refl inside to the-international criminal tribunal

Manual Evaluation of Translation Quality

Divide translations of PVs into:

- *good* - correct translation, correct verb inflection;
- *acceptable* - correct translation, wrong inflection (also when a reflexive particle is missing, or a 'da'-construction is not built correctly);
- *incorrect* - incorrect translation;

Manual Evaluation of Translation Quality

Manual evaluation of translation quality on sentences with correctly identified PVs:

method	translation quality		
	good	acceptable	incorrect
baseline	0.21	0.41	0.39
static	0.25	0.51	0.24
dynamic	0.24	0.51	0.25

Manual Evaluation of Translation Quality (Split Separable PVs)

Manual evaluation of translation quality on sentences with correctly identified separable PVs appearing in a split form (60 occurrences):

method	translation quality		
	good	acceptable	incorrect
baseline	0.05	0.28	0.67
static	0.10	0.62	0.28
dynamic	0.07	0.35	0.58

Manual Evaluation of Translation Quality (Semantic Compositionality)

Manual evaluation of translation quality w.r.t the semantic compositionality of PVs (idiomatic: 'i+'; compositional: 'i-';)

method	translation quality					
	good		acceptable		incorrect	
	i+	i-	i+	i-	i+	i-
baseline	0.10	0.10	0.18	0.23	0.20	0.19
static	0.14	0.11	0.26	0.25	0.08	0.16
dynamic	0.12	0.12	0.25	0.27	0.11	0.14

Evaluation Summary

- Both integration approaches lead to improvements in translation quality
- Separable PVs are handled better with the *static* strategy when in a split form
- Possibility for a targeted approach based on compositionality of PVs: treat idiomatic PVs with the *static*, and compositional PVs with the *dynamic* technique

Conclusion

- Phrasal verbs are challenging for SMT
- Their special treatment brings improvements in translation quality
- Future development
  - Expand CG-3 grammar to handle more constructions (e.g., passives)
  - Increase the amount of *good* translations by including morphological features in the factored translation model