PARAPHRASING BIOMEDICAL SUPPORT VERB CONSTRUCTIONS FOR MACHINE TRANSLATION

Anabela Barreiro

CLUP
CENTRO DE LINGUÍSTICA
UNIVERSIDADE DO PORTO
MAIN GOALS

- Build a body of lexical, syntactic and semantic knowledge around SVC
- Apply this linguistic knowledge to paraphrasing - paraphrase and re-write SVC, such as *fazer uma operação* (*to make an operation/surgery*) with corresponding correct equivalences. The equivalences can be:
  - strong lexical verbs, such as *operar* (*to operate on*)
  - lexical-syntactic and stylistic variants of the original SVC, such as *realizar uma operação* (*to perform an operation/surgery*) or *submeter-se a uma operação/cirurgia* (*to undergo/have an operation/surgery or to be operated on*), depending on the lexical-semantic filling of the arguments and on their argument structure.
- show how these paraphrases increase source text quality
- show how MT results can improve significantly when applying paraphrasing capabilities
GENERAL MOTIVATION

1. MT changed the world of translation
   it can no longer be ignored or underestimated

2. It is increasingly more used and useful
   it often replaces human translation when the client requires gisting

3. However, good quality MT is still an ambitious goal
   after +50 years, MT still eludes researchers and developers who are constantly challenged to provide better translations

How can MT move forward?
CHALLENGES TO MACHINE TRANSLATION

1. homography and common-noun nuance
2. anaphora + distant referential associations
3. ellipsis
4. extra-sentential and extra-textual information + extra-linguistic knowledge
5. lexical divergences, idioms, etc.
6. named entities
7. long sentences and wordiness
8. unusual word order
9. **multiword expressions** (including **support verb constructions**)

2009 NooJ conference and workshop
Anabela Barreiro

Touzeur - Tunisia, 8-10 June 2009
**Support verb construction**

= multiword or complex predicate consisting of a semantically weak verb (the support verb), and a predicate noun, a predicate adjective, or a predicate adverb.

Predicate nouns can be:

- morphologically related to a verb
  
  PT: fazer uma *apresentação* de N = *apresentar* N
  
  EN: pay a *visit* to N = *to visit* N

- autonomous
  
  PT: fazer um *mestrado* - *mestrar*
  
  EN: have *fun* - *to fun*
Why Support Verb Constructions?

- Abundant in language and indispensable to communicate

**ASSESSMENT TEST**: Search pattern in corpora (COMPARA)

\(<\text{dar.V+inf}> + \text{<tomar.V+inf}> + \text{<pôr.V+inf}> + \text{<fazer.V+inf}> + \text{<ter.V+inf}>\)? \(<\text{Modif}>? \langle N\rangle\)

Selected the 1st 100 sentences for each verb and manually annotated the 500 sentences as SVC or NO SVC

<table>
<thead>
<tr>
<th>PT</th>
<th>dar</th>
<th>tomar</th>
<th>pôr</th>
<th>fazer</th>
<th>ter</th>
</tr>
</thead>
<tbody>
<tr>
<td>transl.</td>
<td>give</td>
<td>take</td>
<td>put</td>
<td>make/do</td>
<td>have</td>
</tr>
<tr>
<td>NO SVC</td>
<td>11%</td>
<td>12%</td>
<td>33%</td>
<td>53%</td>
<td>80%</td>
</tr>
<tr>
<td>SVC</td>
<td>89%</td>
<td>88%</td>
<td>77%</td>
<td>47%</td>
<td>20%</td>
</tr>
</tbody>
</table>

**RESULTS**:
- These verbs occur very frequently in SVCs
- Overall, in 64.2% of their occurrence; these verbs are support verbs

- Extensively and systematically studied within the Lexicon-Grammar Theory

[M. Gross and followers], also in Portuguese [Ranchhod, 1990] [Baptista, 2005] [Chacoto, 2005] and in contrastive studies [Salkoff, 1990, 1999]

- Often can be replaced by stylistic variants - paraphrases

- Present several degrees of variability (variable - invariable - idiomatic)

- Semantically weak and ambiguous

- They cannot be translated literally
MT OF SUPPORT VERB CONSTRUCTIONS

Ah... That’s better!

WE PROMPTLY VERIFY THE SUPERIOR QUALITY OF THE RESULTS IN THE SECOND CASE!
Por falta de condições técnicas, ele foi removido para o Hospital das Clínicas, onde se fez uma amputação ao nível do ombro.

<table>
<thead>
<tr>
<th>Translation</th>
<th>Natural Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>FreeTranslation</td>
<td>For lack of technical conditions, he was removed for the Hospital of the Clinics, where was done an amputation in terms of shoulder.</td>
</tr>
<tr>
<td>WorldLingo</td>
<td>Due to conditions techniques, it it was removed for the Hospital of the Clinics, where if the shoulder level made an amputation.</td>
</tr>
</tbody>
</table>

Por falta de condições técnicas, ele foi transportado para o Hospital das Clínicas, onde os médicos amputaram o seu braço ao nível do ombro.

<table>
<thead>
<tr>
<th>Translation</th>
<th>Natural Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>FreeTranslation</td>
<td>Due to conditions techniques, it it was carried to the Hospital of the Clinics, where the doctors had amputated its arm to the level of the shoulder.</td>
</tr>
<tr>
<td>WorldLingo</td>
<td>For lack of technical conditions, he was transported for the Hospital of the Clinics, where the doctors amputated his arm level with the shoulder.</td>
</tr>
</tbody>
</table>

Por falta de condições técnicas, ele foi transportado para o Hospital das Clínicas, onde o seu braço foi amputado ao nível do ombro.

<table>
<thead>
<tr>
<th>Translation</th>
<th>Natural Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>FreeTranslation</td>
<td>for lack of technical conditions, he was transported for the Hospital of the Clinics, where arm was amputated level with the shoulder.</td>
</tr>
<tr>
<td>WorldLingo</td>
<td>Due to conditions techniques, it it was carried to the Hospital of the Clinics, where its arm was amputated to the level of the shoulder.</td>
</tr>
</tbody>
</table>
Linguistic knowledge of paraphrases can improve the quality of machine translation

- The conversion of support verb constructions into morphosyntactically and/or semantically related verbs or verbal expressions and the adequate specification and inclusion of predicate-argument structure information produces a controlled language that is applicable both to general and domain specific contexts and makes MT more reliable

1. Reduces complexity (= paraphrasing by simplification)
2. Reduces lexical ambiguity (paraphrase conveys the correct interpretation)
3. Facilitates interpretation when the support verb construction is idiomatic
4. Reduces wordiness (sometimes, improves source text quality)
5. Improves translation quality
THEORETICAL BACKGROUND - LINGUISTICS


- **FrameNet** [Fillmore et al., 2002, 2003] – records the information necessary for the representation of argument mapping relations between a support verb and a nominalization.

- **NomBank** [Meyers et al., 2004b, 2004b] - maps syntactic positions in nominalizations to verbal arguments and identify the allowed complements for a nominalization, relating the nominal complements to the arguments of the corresponding verb, including information about support verbs.

- **Lexicon-Grammar Theory** [M. Gross and followers] - theoretical framework adopted in this research
Related work on paraphrasing

- **question answering** – discovering paraphrased answers provides additional evidence that an answer is correct - [Ibrahim et al., 2003], [Pașca, 2003], [Duboué & Chu-Carroll, 2006]

- **information extraction** and **text mining** - paraphrases help text categorization tasks or mapping to texts with similar characteristics - [Ibrahim et al., 2003], [Shinyama et al., 2002] [Shinyama & Sekine, 2003], [Sekine, 2005] [Pașca, 2005], [Pașca & Dienes, 2005]

- **summarization** - the identification of paraphrases allows information across documents to be condensed and helps improve the quality of the generated summaries [McKeown et al., 2002], [Barzilay, 2001, 2003], [Hirao et al., 2004] [Zhou et al., 2006b]

- **natural language generation** – the generation of paraphrases allows the production of more varied and fluent text - [Iordanskaja et al. 1991]

- **machine translation** – paraphrases help create a more fluent translation and are valuable in the evaluation of MT results - [Zhou et al., 2006], [Callison-Burch et al., 2006a, 2006b, 2007 and 2008]
Related work on Controlled Writing and Stylistic Editors/Authoring Aids

- **Controlled language** - application of human language techniques in an industrial environment. It is used to support technical writers in producing high quality technical documentation by checking spelling, grammar, style and terminology in technical documents (text optimisation). Automated rewriting for controlled language translation.

- **Stylistic editing** - focuses on clarity and expression, re-ordering of paragraphs, sentences and words; elimination of verbiage and jargon; simplification of untangle complicated clauses; elimination of inconsistencies, use of stylistic paraphrases (reductions/extensions).

- **Authoring aids** - contemplate mostly spell checking, simple grammar checking (correction of text), writing and morphological variants and synonyms LREC 2008

Some available tools:

MULTIDOC [Haller, 2000]

KANT CE Checker [Mitamura and Nyberg, 2001] [Mitamura et al., 2003] [Rascu, 2006]

CLAT [Schmidt-Wigger, 1998]; [Carl et al., 2002]; [Hernandez and Rascu, 2004]

CLOUT - the Controlled Language Optimized for Machine Translation.
**METHODOLOGY**

Task 1

Analysis, extraction from **corpora** and formalization of support verb constructions by using information described in **dictionaries** and **grammars** – NooJ linguistic environment was the tool used

Task 2

**Paraphrasing** of predicate noun constructions

Phase 1

Development of **resources** and **pre-processing**

Construction of **simple grammars** to recognize and extract SVCs from texts

Definition of **properties in the dictionary** and establishment of **semantic links** between verbs and SVCs + establishment of classes of paraphrases

Application of grammars with dictionary information to texts

Refinement of grammars

**Finer recognition** of SVCs in texts -> **paraphrasing** + translation

Phase 2

**Evaluation** experiments
Explicit marking of derivation and semantic verb associations

Adjective entries:
• Identification of derivational paradigms for predicate adverbs - adverbializations (annotation AVDRV)

Autonomous predicate nouns:
• Identification of autonomous predicate nouns (annotation Npred)
• Identification of a semantically related verb (annotation VRB)
• Link to the support verb (annotation VSUP)

curso,N+FLX=ANO+Npred+IN+inst+EN=course+VSUP=tirar+VRB=estudar+NPrep=de+Det=um
Explicit marking of derivation and support verb associations

Verb entries:

• Identification of derivational paradigms for predicate nouns - nominalizations (annotation $NDRV$) and predicate adjectives (annotation $ADRV$)

• Link to the support verbs (annotation $VSUP$) occurring with the derived predicate nouns and adjectives

adaptar,V+FLX=FALAR+Aux=1+INOP57+Subset=132+EN=adapt+$VSUP=fazer$+$DRV=NDRV00$:CANÇÃO
azedar,V+FLX=LIMPAR+Aux=1+OBJTRundif98+Subset=740+EN=sour+$VSUP=estar$+$DRV=ADRV00$:ALTO
NEW RESOURCES

Port4NooJ

• a set of open source Portuguese linguistic resources representing ontological relations, and paraphrasal relations. This resources integrates a bilingual extension for PT-EN MT.

<table>
<thead>
<tr>
<th>Semantic relations</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Verb - Predicate Nouns</td>
<td>8472</td>
</tr>
<tr>
<td>Predicate Adjectives - Adverbs</td>
<td>222</td>
</tr>
<tr>
<td><strong>Total Links</strong></td>
<td><strong>8694</strong></td>
</tr>
</tbody>
</table>

• Publicly available resources at:

http://www.nooj4nlp.net
http://www.linguateca.pt/Repositorio/Port4Nooj/

DicTUM (Diccionário de Termos e Unidades Multipalavra)

• a Dictionary of Multiword Expressions
NEW TOOLS

➢ **ReWriter**
  - a monolingual standalone paraphraser to pre-edit texts, using paraphrasing capabilities
  - Portuguese version **ReEscreve** - publicly available service at: http://www.linguateca.pt/ReEscreve/

➢ **ParaMT**
  - a bilingual/multilingual paraphraser to be integrated in machine translation systems
USEFULNESS OF THE AFOREMENTIONED PARAPHRASING TOOLS

- Controlled writing & text pre-editing
- Machine translation
- Text production and stylistics (authoring aids)
Anabela Barreiro
Touzeur - Tunisia, 8-10 June 2009
ParaMT: A Bilingual/Multilingual Paraphraser for MT

Recognition of Portuguese SVC and translation into English verbs

Machine translation

2009 NooJ conference and workshop
Anabela Barreiro

Touzeur - Tunisia, 8-10 June 2009
**Evaluation experiment 1**

5 support verbs
500 baseline sentences - manually annotated
100 for each elementary support verb

<table>
<thead>
<tr>
<th></th>
<th>SVC Recognition Precision</th>
<th>SVC Recognition Recall</th>
<th>SVC Paraphrasing Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pôr</td>
<td>73/73 - 100%</td>
<td>73/100 – 73%</td>
<td>72/73 - 98.6%</td>
</tr>
<tr>
<td>Tomar</td>
<td>75/75 - 100%</td>
<td>75/100 – 75%</td>
<td>68/73 - 93.1%</td>
</tr>
<tr>
<td>Ter</td>
<td>65/65 - 100%</td>
<td>65/100 – 65%</td>
<td>59/65 - 90.7%</td>
</tr>
<tr>
<td>Dar</td>
<td>57/60 - 95%</td>
<td>57/100 – 57%</td>
<td>46/51 - 90.1%</td>
</tr>
<tr>
<td>Fazer</td>
<td>43/45 – 95.5%</td>
<td>43/100 – 43%</td>
<td>40/45 - 88.8%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>98.4%</strong></td>
<td><strong>62.6%</strong></td>
<td><strong>93.4%</strong></td>
</tr>
</tbody>
</table>
**Evaluation Experiment 2**

5 different support verbs

100 sentences with SVC (for each language PT and EN)

20 sentences for each support verb – manually assigned a paraphrase for each one of them
> 20 SVC sentences + 20 paraphrases of the original SVC sentences

TOTAL: 200 sentences for each language
100 with SVC + 100 with paraphrases of the SVC

10 testers translated and evaluated translations
students of the MA in Translation and Linguistic Services

<table>
<thead>
<tr>
<th></th>
<th>Equal quality result</th>
<th>SVC translation is better</th>
<th>Paraphrase translation is better</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>EN-PT MT</strong></td>
<td>17%</td>
<td>26%</td>
<td>57%</td>
</tr>
<tr>
<td><strong>PT-EN MT</strong></td>
<td>19%</td>
<td>30%</td>
<td>51%</td>
</tr>
</tbody>
</table>

For both EN-PT and PT-EN, testers considered that the paraphrase translations presented a better quality. The evaluation results clearly confirm that these paraphrases helped the systems produce better MT.
### Enhanced Dictionary Entries with Syntactic-Semantic Information

<table>
<thead>
<tr>
<th>Word</th>
<th>Gloss</th>
<th>Syntactic-Semantic Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>adaptar</td>
<td>V+FLX=FALAR+Aux=1+INOP57+Subset132+EN=adapt+VSUP=fazer+DRV=NDRV00:CANÇÃO+NPrep=de</td>
<td>favor,N+FLX=MAR+Npred+AB+state+EN=favor+VSUP=fazer+NPrep=a+VRB=ajudar</td>
</tr>
<tr>
<td>amputar</td>
<td>V+FLX=FALAR+Aux=1+OBJTRundif21+BioMed+EN=amputate+SUJ=AG+VSUP=fazer+DRV=NDRV00:CANÇÃO+NPrep=de+OD=BP+VSTYLE=realizar+VSTYLE=efectuar+VASP=iniciar+VASP=prosseguir+VASP=concluir</td>
<td></td>
</tr>
<tr>
<td>rápido</td>
<td>A+FLX=RÁPIDO+PV=eagerType+EN=quick+DRV=AVDRV06:RAPIDAMENTE</td>
<td></td>
</tr>
<tr>
<td>adocar</td>
<td>V+FLX=COMEÇAR+Aux=1+OBJTRundif75+Subset604+EN=sweeten+DRV=ADRV11:VERDE+VCOP=tornar</td>
<td></td>
</tr>
<tr>
<td>transplantar</td>
<td>V+FLX=FALAR+Aux=1+RECTR26+Subset=504+BioMed+EN=transplant+SUBJ=AG+VSUP=fazer+DRV=NDRV79:ANO+NPrep=de+DO=BP+VSTYLE=realizar+VSTYLE=efectuar+VASP=iniciar+VASP=prosseguir+VASP=concluir</td>
<td></td>
</tr>
<tr>
<td>médico</td>
<td>N+FLX=ANO+AN+med+AG+EN=doctor</td>
<td></td>
</tr>
<tr>
<td>médico</td>
<td>N+FLX=ANO+AN+med+AG+EN=physician</td>
<td></td>
</tr>
<tr>
<td>doente</td>
<td>N+FLX=ANO+AN+med+PAT+EN=patient</td>
<td></td>
</tr>
</tbody>
</table>
GRAMMAR TO RECOGNIZE AND PARAPHRASE BIOMEDICAL SVC
PARAPHRASING FOR CONTROLLED WRITING AND TEXT STYLISTICS

ARG0 = AGENT (AG)
Elementary SVC > Lexical Verb – fazer uma amputação = amputar
(to amputate)
Elementary SVC > non-elementary SVC - realizar/efectuar uma amputação
(to perform an amputation)

ARG0 = PATIENT (PAT)
Submeter-se/ser submetido a uma operação (to undergo surgery)
Ser operado (to be operated)
Elementary SVC > non-elementary SVC - realizar/efectuar uma operação
(to perform an operation)
RECOGNITION AND MONOLINGUAL PARAPHRASING OF BIOMEDICAL-RELATED SVC

nça, o cirurgião Faivre, ao
nça, o cirurgião Faivre, ao
nça, o cirurgião Faivre, ao
1 ser interrogadas antes de
1 ser interrogadas antes de
1 ser interrogadas antes de
1 ser interrogadas antes de
o público de saúde recusa
o público de saúde recusa
Tiago Felizardo, vai ter de
Tiago Felizardo, vai ter de
Tiago Felizardo, vai ter de
Tiago Felizardo, vai ter de
aber se o doente consegue
aber se o doente consegue
aber se o doente consegue
aber se o doente consegue
médico também lhe pode
médico também lhe pode
médico sempre vai querer
médico sempre vai querer
britânico, conseguiu
britânico, conseguiu
os pacientes que precisam
os pacientes que precisam
os pacientes que precisam
os pacientes que precisam
fazer uma amputação/amputar
fazer uma amputação/efectuar uma amputação
fazer uma amputação/realizar uma amputação
fazer um aborto/submeter-se a um aborto
fazer um aborto/abortar
fazer um aborto/efectuar um aborto
fazer um aborto/realizar um aborto
fazer uma operação cirúrgica/realizar uma operação cirúrgica
fazer uma operação cirúrgica/efectuar uma operação cirúrgica
fazer uma operação plástica depois de/sujeitar-se a uma operação plástica
fazer uma operação plástica depois de/submeter-se a uma operação plástica
fazer uma operação plástica depois de/realizar uma operação plástica
fazer uma operação plástica depois de/efectuar uma operação plástica
fazer uma prova de esforço/sujeitar-se a uma prova de esforço
fazer uma prova de esforço/submeter-se a uma prova de esforço
fazer uma prova de esforço/realizar uma prova de esforço
fazer uma prova de esforço/efectuar uma prova de esforço
fazer uma prova de esforço para/realizar uma prova de esforço
fazer uma prova de esforço para/efectuar uma prova de esforço
fazer um transplante/de/realizar um transplante
fazer um transplante/de/efectuar um transplante
fazer uma transfusão de sangue/realizar uma transfusão de sangue
fazer uma transfusão de sangue/efectuar uma transfusão de sangue
fazer uma transfusão de sangue/sujeitar-se a uma transfusão de sangue
fazer uma transfusão de sangue/submeter-se a uma transfusão de sangue
fazer uma transfusão de sangue/realizar uma transfusão de sangue
fazer uma transfusão de sangue/efectuar uma transfusão de sangue
MAIN CONCLUSION

- Linguistic knowledge on paraphrases formalized in this research when applied to a MT system improves its output quality.

When SVC were identified and replaced with semantically equivalent or similar verbal expressions as a pre-processing step to translating:

- a **21% improvement** was observed in the evaluated quality of the results of PT-EN machine translation
- a **31% improvement** in the results of EN-PT machine translation.
FUTURE RESEARCH SUGGESTED BY THIS THESIS

- Enlargement of syntactic-semantic relations between predicates, and other morphossyntactically and/or semantically related elements
  
  Exs: *operar* – *operação* – *operado*
  *urgente* – *urgentemente*

- Assignment of thematic roles and semantic functions to nominal entities
  
  Exs: *médico*+AG+EN=*doctor*  
  *doente*+PAT+EN=*patient*

- Paraphrasing of various linguistic phenomena for ReWriter extensibility (comprising stylistic variance and controlled language)

- Development and enhancement of ParaMT

- Enhancement of the machine translation model
<table>
<thead>
<tr>
<th>Linguistic Phenomenon</th>
<th>Expression or sentence</th>
<th>Paraphrase</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Adverbial</strong></td>
<td>à volta da órbita</td>
<td>periorbital</td>
</tr>
<tr>
<td></td>
<td><em>around the orbit of the eye</em></td>
<td><em>periorbital</em></td>
</tr>
<tr>
<td></td>
<td>de forma interactiva</td>
<td>interactivamente</td>
</tr>
<tr>
<td></td>
<td><em>in an interactive way</em></td>
<td><em>interactively</em></td>
</tr>
<tr>
<td><strong>Relative Clause</strong></td>
<td>N0 que têm sido escritos</td>
<td>N0 que foram escritos</td>
</tr>
<tr>
<td></td>
<td><em>N0 that have been written</em></td>
<td><em>N0 that were written</em></td>
</tr>
<tr>
<td></td>
<td>A velocidade a que se move a luz</td>
<td>A velocidade da luz</td>
</tr>
<tr>
<td></td>
<td><em>The speed to which light moves</em></td>
<td><em>The speed of light</em></td>
</tr>
<tr>
<td></td>
<td>O papel que a Europa tem</td>
<td>O papel da Europa</td>
</tr>
<tr>
<td></td>
<td><em>The role that Europe plays/has</em></td>
<td><em>The role of Europe ≡ Europe’s role</em></td>
</tr>
<tr>
<td></td>
<td>As dificuldades que temos</td>
<td>As nossas dificuldades</td>
</tr>
<tr>
<td></td>
<td><em>The difficulties we have</em></td>
<td><em>Our difficulties</em></td>
</tr>
<tr>
<td><strong>If clause</strong></td>
<td>se for necessário</td>
<td>se necessário</td>
</tr>
<tr>
<td></td>
<td><em>if it is necessary</em></td>
<td><em>if necessary</em></td>
</tr>
<tr>
<td><strong>Named Entity</strong></td>
<td>A rainha de Inglaterra</td>
<td>A rainha inglesa</td>
</tr>
<tr>
<td></td>
<td><em>The queen of England</em></td>
<td><em>The British queen</em></td>
</tr>
<tr>
<td><strong>Noun Phrase</strong></td>
<td>O heróico povo português</td>
<td>Os heróicos portugueses</td>
</tr>
<tr>
<td></td>
<td><em>The heroic Portuguese people</em></td>
<td><em>The heroic Portuguese</em></td>
</tr>
</tbody>
</table>

A formal linguistic study of paraphrases as these ones would represent a significant contribution to NLP in general, and to MT in particular.
RELEVANT PUBLICATIONS


  http://poloclup.linguateca.pt/Port4NooJ/ResourcesOverview.pdf


REFERENCES


ACKNOWLEDGMENTS

This research work was funded by Fundação para a Ciência e a Tecnologia (doctoral scholarship SFRH/BD/14076/2003), co-financed by POSI and partly supported by Fundação para a Computação Científica Nacional - Linguateca.

In addition to the doctoral scholarship, I had a ½ month contract from Linguateca to help make Port4NooJ and ReEscreve public.

I currently have a new contract with Linguateca to enlarge and enhance the linguistic resources herein described.

شكرا جزيلًا