Extraction of idioms from dictionaries seen as specialized corpora: ≈20000 PLEs in por-fra

Christian BOITET & Mathieu MANGEOT
GETALP-LIG, Grenoble, France
Christian.Boitet@imag.fr, Mathieu.Mangeot@imag.fr
AIM-West meeting, 3-4/10/2016

Terms

- MWEs ? Rather PLEs (polylexical expressions)
  By definition, a MWE is composed of several [segmentable] words
  - Even in French, that is not clear:
    1. plateforme — plateformes
    2. plate-forme — plates-formes
    3. compte rendu — comptes rendus
    4. pomme de terre — pommes de terre
    5. prendre $ son temps — ils prennent $=très souvent leur temps
  - Compound words often cannot be segmented in usual words
    1. Arbeitszeit — NOT Arbeits Zeit
    2. Antimoine — NOT anti moine
    3. Wiedersehen — NOT wieder sehen
- Traditionally, MT developers spoke of

  idioms — *tournures*

  idiomatic expressions — *expressions idiomatiques*

Outline

- Overall motivation and context
  Build HQ large-coverage MT systems
  The quality of a MT system is that of its idiom dictionary [N.Nédobejkine 1970]
  Not only quality, but also quantity is important for "large coverage"
- Idiom handling in large-coverage MT systems
  - Systran (since 1967),
  - Ariane/RUS-FRA, FRA-ENG (since 1980), ETAP-3 (since 1985)
  - METAL (since 1984), Word Magic ENG-SPA (ConverterForHealthcare)
  - Many Japanese systems (at least for terminology)

  Types of idioms more or less well handled
  - fixed and connex — parce que
  - variable and connex — pomme de terre / pommes de terre
    compte rendu / comptes rendus
  - not connex — he gave a lot of money back to the customer
    er hörte nach 5 Minuten auf, Lärm zu machen

Problem statement & possible angles of attack

- How to get a large quantity of HQ pairs of idioms in L1-L2?

  Human work (lexicographers)
  - Not so expensive (e.g. Systran used retirees and college dropouts)
  - Sources
    1. There are many for terminology (e.g. regular publications in Russia, USA…)
    2. There are quite rare and incomplete in usage dictionaries

  Automatic extraction from parallel corpora followed by human cleaning
  - Usable for simple as well as compound lexemes
  - Ex: ATLAS-II (Fujitsu), a pivot-based MT system
    1. MT-summit 1001: 586000 entries (jp-concepts-en)
    2. ACL-2003: 1M entries
    3. COLING-2004: 3M entries
    4. 2009: 6.5 M entries
  - Not so good for non-terminological or inflected or non-connex idioms probably because
    1. they are much less frequent
    2. they are more difficult to identify (inflected forms, discontinuousness, agreement)

Possible angles of attack (cont.) + our approach

Extraction from monolingual corpora followed by translation

- not very productive, and noisy even if using a parser (cf. V. Seretan)
- translation still very difficult and costly
  1. post-editing of MT output on HQ source idioms ? Circularity
  2. direct human translation? Why not, but supposes a very high level in L1 & L2

- A second problem: how to attach MT-useful information?

Type of idiom

- terminological
- LSF-based collocation like support verb…

Morphological and syntactic variability

- possibility of passivation?
- possibility of inserting words / syntagms?
- realization of variable parts (e.g. to have one's say, she has her say)

- Our approach to the first problem (extraction)

Use automatic extraction from "special corpora"

- made of documents containing many idiom pairs of HQ
- easily parsable

Hence, we looked at collections of bilingual dictionaries in e-book form

An experiment to get idioms in Portuguese-French

- Choice of corpus (dictionary collection)
  - Visit to bookshops and manual inspection of dictionaries
  - Selection of a the Porta Editora collection
    - quite large size (authors say they put “everything they could”)
    - a prima vista, good quality
    - very simple presentation of idiom lists
  - At this moment, por-fra, fra-por, por-eng, spa-por
- A deceptively simple process
  - Purchase of paper and e-book forms of selected dictionaries
    - a few days and ≈2*20€/book
  - Extraction (from the e-book form) of the underlying docx format
    - 15 minutes
  - Load in TextWrangler and use 2 or 3 regular expressions to get result
    - 15 minutes

Excerpt from tournures-por-fra.txt (961 K)

```
ele zangou-se com o melhor amigo <=> il s'est brouillé avec son meilleur ami
trabalhar com zelo <=> travailler avec zèle
jardim zen <=> jardin zen
cinco graus abaixo de zero <=> cinq degrés au-dessus de zéro
dos zero aos dez anos <=> de zéro à dix ans
zero a zero <=> zéro à zéro
ganhar por três a zero <=> gagner par trois à zéro
merecer um zero <=> mériter un zéro pointé
zero absoluto <=> zéro absolu
partir do zero <=> partir de zéro
ser um zero à esquerda <=> être un zéro
voltar à estaca zero <=> retourner à la case départ
ir aos ziguezagues <=> zigzaguer
óxido de zircónio <=> zircone
zona de alta/baixa pressão <=> zone de haute/basse pression
zona de operações <=> zone des opérations
zona glacial <=> zone glaciale
zona industrial <=> zone industrielle
zona neutra <=> zone neutre
zona ocupada <=> zone occupée
zona temperada <=> zone tempérée
zona tórrida <=> zone torride
zona verde <=> espace vert
sentir-se zonzo <=> être pris de vertiges
```

Current result and planned future work

- Current result
  - obtained from por-fra part only
  - 20000 idiom pairs extracted
  - no annotation (nothing on type, variability, frequency, domain…)
- Planned future work on problem 1 (primary extraction)
  - do the same with the fra-por volume
  - merge (and check)
  - do the same for por-eng and spa-eng
  - put them in the PIVAX-3 lexical database
  - using the English volumes, build a UNL volume for these PLEs
    1. we can use special notation available for compound expressions in UNL
    2. but we have to derive some similar notation from that and Parseme scheme

Perspectives to attack problem 2

- Problem 2 is about finding and coding information useful for MT

  Type of idiom
  - terminological (connex but allowing inflections), + domain
  - nominal non-compositional expression (compte rendu — minutes)
  - LSF-based collocation like support verb construction
  - others, with subtypes (e.g. “avoir son mot à dire”)

  Morphological and syntactic variability
  - possibility of passivation?
  - possibility of inserting words / syntagms?
  - constraints on the realization of variable parts
    e.g. to have one’s say, she has her say
- Possible approach
  - Look for (automatically generated) possible instances of these PLEs in monolingual texts
  - Identify the “variability type” of each PLE
  - Try to find as yet “undiscovered” PLEs, using the known ones as sign posts

What about MT?

- Next step is to integrate these ≈20000 idioms in an MT system
- Possibilities

  PAHOMTS?
  - A cooperation was envisaged with PAHO through Ms De La Fuente long ago but it never materialized

  Systran?
  - Maybe, but no cooperation projects with them succeded until now

  Paltonio Daun Fraga's Ariane-based MT system
  - Yes, but he redeveloped and improved only his POR-ENG system
    1. That was last year, for AIM-West, using the Ariane-Heloise platform (now available on the lingwarium.org cooperative website)
    2. A way to produce minimal annotations should be found
  - It should be possible to develop a POR-FRA system in 2017 by connecting to 1 of the Ariane-based French generation modules
- Why not embark on a common MT project centering on HQ PLEs?

  Languages: POR, FRA, ENG, DEU, RUS (coop. with ETAP-3), & UNL

  Possible by-product: active reading tool showing corresponding PLEs