Per-Fide: Latest Developments

Rui Brito

Universidade do Minho

Braga, Setembro 2013

Index

1. Goals (Recap)
2. TMXA (Recap)
3. PTD and PTDA
4. Workflow
5. Conclusions and future work

Goals (Recap)

- Generate annotated TMXs (TMXA);
  - Other formats are too complex (TEI) or documentation is non-existent (XCES);
  - The TMX format is simple, we just need to add some more information to it;
- Generate partially annotated PTDs from TMXA;
- Sharing annotated documents;
- Broaden the project's scope;
- Robustness;

TMXA (Recap)

- A kind of merge between CWB and TMX formats;
- Before tagging, the TMX is atomized;
- TMX allows for notes in the header;
  - TMXA may be a list of attributes by column in the header;
- Entries separated by newlines;
- Each line contains the word, followed by lemma and ending with POS;
  - CWB format;
  - Economic (space-wise);
- Allows searches by lemma;

Listing 1: TMX example

```
<tuv lang=’en’><seg>And much poison at last for a pleasant death .<
    /seg></tuv>
```

Listing 2: TMXA example

```
<tuv lang=’en’><seg><![CDATA[<s>
And     And     <cnjcoo>
much    much    <det><qnt><sg>
poison  poison  <n><sg>
at      at      <pr>
last    last    <det><ord><sp>
for     for     <pr>
a       a       <det><ind><sg>
pleasant        pleasant        <adj>
death   death   <n><sg>
.       .       <sent>
</s>
]]></seg></tuv>
```

PTD and PTDA

- Generated from TMXA;
- Partially annotated;
  - Reduces noise due to certain linguistic constructs;
  - With less linguistic forms, the algorithm converges much faster;

Original PTD:

```
"great" => {
  count => 31,
  trans => {
           "grande" => 0.64500004,
          "grandes" => 0.08481482,
            "muito" => 0.05734283,
        "destinado" => 0.03269691,
           "pesado" => 0.03253261,
                "A" => 0.03033091,
         "avançava" => 0.01189576,
      "percorremos" => 0.00946403,
    },
},
```

PTD after processing:

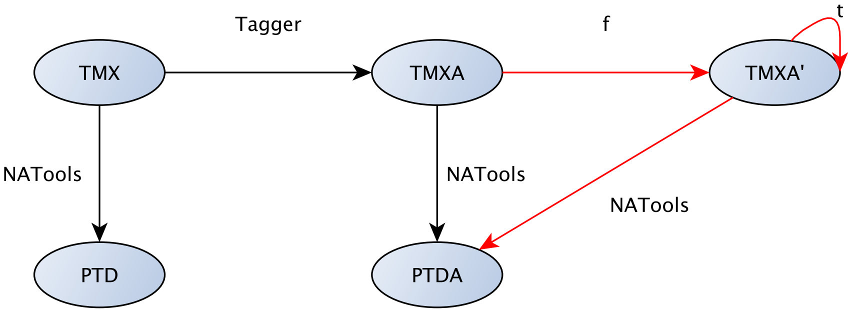
```
"_great" => {
  count => 46,
  trans => {
          "_grande" => 0.51510072,
           "_maior" => 0.27329040,
            "muito" => 0.06289405,
          "_sentir" => 0.02665226,
     "concordância" => 0.02204488,
          "afectos" => 0.02167012,
            "ardor" => 0.02138354,
            "fardo" => 0.02052378,
    },
},
```

Original PTD:

```
"imaginar" => {
  count => 118,
  trans => {
          "imagine" => 0.57757628,
           "(none)" => 0.03993374,
        "imagining" => 0.03647405,
           "fathom" => 0.03638203,
         "wondered" => 0.03180898,
          "picture" => 0.02748435,
         "imagined" => 0.02548767,
         "conceive" => 0.01833824,
    },
```

PTD after processing:

```
"v_imaginar" => {
  count => 226,
  trans => {
        "v_imagine" => 0.46072519,
         "v_wonder" => 0.11047629,
          "v_think" => 0.05020738,
         "pictured" => 0.03797891,
           "(none)" => 0.03407935,
       "pp_imagine" => 0.02771692,
        "pp_wonder" => 0.02708688,
           "fathom" => 0.02363599,
    },
},
```


Tagger
f
t
TMX
TMXA
TMXA'
NATools
NATools
NATools
PTD
PTDA

Conclusions and future work

- Versatile format;
- Tools that can handle TMX, can also handle TMXA;
- Handling of difficult elements was further improved;
  - Contractions;
  - Multiple word elements;
- Development of an easy to use abstraction layer;
- Further improve treatment of multiple word elements;
- Tagset still need to be worked on:
  - Apertium tagset fully supported;
  - Freeling (parole) tagset still needs work, but its growing;

Per-Fide: Latest Developments

Rui Brito

Universidade do Minho

Braga, Setembro 2013

– ? –