GROBID-DICTIONARIES: INFRASTRUCTURE FOR AUTOMATICALLY STRUCTURING DIGITIZED DICTIONARIES AND ENTRY-BASED DOCUMENTS

Mohamed Khemakhem
mohamed.khemakhem@inria.fr
Standard-based Lexical Models for Automatically Structured Dictionaries
Context: PhD Thesis

Standard-based Lexical Models for Automatically Structured Dictionaries
Dictionaries

- **Digitisation**
  - Digitally Born
  - Digitised

- **Content**
  - Lexical
  - Encyclopedic

- **Time**
  - Legacy
  - Modern
Automatically Structuring: Landscape

- Domination of rule based methods

- Just 2 previous works used machine learning
  - Crist, S. (2011)
Standard-based Lexical Models

- Standardized resources are required for large scale projects
  - Enable mapping and enriching

- Text Encoding Initiative (TEI)
  - Too flexible

- Lexical Markup Framework (ISO 24613 LMF)
  - Too restricted
GeneRation Of Bibliographic Data = GROBID
GROBID: GeneRation Of Bibliographic Data

- Java library for processing and structuring text information (mainly in PDF)

- Supervised machine learning for Information Extraction
  - Sequence Labelling: Conditional Random Fields (CRF)

- TEI-compliant output

- Used by: Research Gate, HAL, Mendeley, CERN,...
Depth-resolved analysis of spontaneous phase separation in the growth of lattice-matched AlInN

A. Redondo-Cubero¹,²*, K. Lorenz², R. Gago⁴, N. Franco⁵, M.-A. di Forte Poisson², E. Alves¹ and E. Muñoz³

¹ISOM and Dpt. de Ingeniería Electrónica, ETSI Telecomunicación, Universidad Politécnica de Madrid, E-28040 Madrid, Spain.
²Centro de Micro-Análisis de Materiales, Universidad Autónoma de Madrid, E-28049 Madrid, Spain.
³Instituto Tecnológico e Nuclear, Estrada nacional 10, 2686-953 Sacavém, Portugal.
⁴Instituto de Ciencia de Materiales de Madrid (CSIC), E-28049 Madrid, Spain.
⁵Thales Research & Technology/TIGER, 91461 Marcoussis Cedex, France.

ABSTRACT:

We report the detection of phase separation of an Al₁ₓInₓN/GaN heterojunction grown close to lattice matched conditions (x~0.18) by means of Rutherford backscattering spectrometry in channeling geometry and high resolution x-ray diffraction. An initial pseudomorphic growth of the film was found, with good single crystalline quality, the
GROBID-Family

GeneRation Of Bibliographic Data = GROBID

- GROBID-Dictionaries
- GROBID-NER (Entity Fishing)
- GROBID-Quantities
- GROBID-Astro
- GROBID-Dictionaries

GROBID-NER (Entity Fishing)
GROBID-Dictionaries: Approach
condenser [kə'dəsə] v. t. (lat. condensare, rendre épais). Rendre plus dense, réduire à un moindre volume. || Liquéfier un gaz par refroidissement ou compression : le froid condense la vapeur d'eau. || Fig. Exprimer d'une manière concise, en peu de mots :
<table>
<thead>
<tr>
<th>CRF model</th>
<th>Dictionary Segmentation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Segmented Page</td>
<td>headnote  body  footnote  &lt;dictScrap&gt;</td>
</tr>
<tr>
<td>CRF model</td>
<td>Dictionary Body Segmentation</td>
</tr>
<tr>
<td>Segmented Page</td>
<td>&lt;entry&gt;  &lt;dictScrap&gt;</td>
</tr>
<tr>
<td>CRF model</td>
<td>Lexical Entry</td>
</tr>
<tr>
<td>Segmented LE</td>
<td>&lt;form&gt;  &lt;etym&gt;  &lt;sense&gt;  &lt;re&gt;  &lt;dictScrap&gt;</td>
</tr>
<tr>
<td>CRF models</td>
<td>Form  Sense  Grammatical Group</td>
</tr>
<tr>
<td>Segmented Form/Sense</td>
<td>&lt;orth&gt;  &lt;pron&gt;  &lt;gramGrp&gt;  &lt;sense&gt;  &lt;dictScrap&gt;</td>
</tr>
<tr>
<td>CRF model</td>
<td></td>
</tr>
</tbody>
</table>

**GROBID-Dictionaries: Architecture**

- **Dictionary Segmentation**
- **Dictionary Body Segmentation**
- **Lexical Entry**
- **Form**
- **Sense**
- **Grammatical Group**
GROBID-Dictionaries: Output Serialisation

- **TEI-Lex-0**
  - New recommendations for unified TEI modelling

- **LMF**
  - Review the meta model
  - Propose a new TEI-based serialisation: ISO 24613-4
CRF Models: Feature Selection
CRF Models: Feature Selection
GROBID-Dictionaries: Enhanced Usability

Khemakhem et al. 2018a
Use cases: Lexical Dictionaries

- Khemakhem et al. 2017
  - Monolingual digital English dictionary
  - Bilingual digitised Fang-French dictionary

<table>
<thead>
<tr>
<th>label</th>
<th>accuracy</th>
<th>precision</th>
<th>recall</th>
<th>f1</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;form&gt;</td>
<td>99.69</td>
<td>99.31</td>
<td>97.28</td>
<td>98.28</td>
</tr>
<tr>
<td>&lt;pc&gt;</td>
<td>99.39</td>
<td>100</td>
<td>78.79</td>
<td>88.14</td>
</tr>
<tr>
<td>&lt;re&gt;</td>
<td>98.49</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>&lt;sense&gt;</td>
<td>88.57</td>
<td>85.18</td>
<td>98.64</td>
<td>91.42</td>
</tr>
</tbody>
</table>

Evaluation of 'Lexical Entry" model on EEBD

<table>
<thead>
<tr>
<th>label</th>
<th>accuracy</th>
<th>precision</th>
<th>recall</th>
<th>f1</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;form&gt;</td>
<td>90.77</td>
<td>57.94</td>
<td>75.26</td>
<td>65.47</td>
</tr>
<tr>
<td>&lt;pc&gt;</td>
<td>97.6</td>
<td>28.57</td>
<td>11.76</td>
<td>16.67</td>
</tr>
<tr>
<td>&lt;sense&gt;</td>
<td>92.45</td>
<td>96.46</td>
<td>93.59</td>
<td>95</td>
</tr>
</tbody>
</table>

Evaluation of 'Lexical Entry" model on FFD
Use cases: Encyclopedic Dictionaries and MAC

- Khemakhem et al. 2018b

Source: Manuscript Auction Catalogues (1889)

Source: Petit Larousse Illustré (1906)
Use cases: Legacy Address Directories

- Khemakhem et al. 2018c

Source:
https://gallica.bnf.fr/ark:/12148/cb32695639f/date19010101
Use cases: Legacy Address Directories

- Khemakhem et al. 2018c

<table>
<thead>
<tr>
<th>Model</th>
<th>Annotated Data</th>
<th>F1-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dictionary Segmentation</td>
<td>10 Pages, 7 training, 3 evaluation</td>
<td>Micro Average</td>
</tr>
<tr>
<td>Dictionary Body Segmentation</td>
<td>319 Entries, 270 training, 49 evaluation</td>
<td>Micro Average</td>
</tr>
<tr>
<td>Lexical Entry</td>
<td>208 Entries, 160 training, 48 evaluation</td>
<td>Micro Average</td>
</tr>
</tbody>
</table>

Evaluation of the first three segmentation models

Source: https://gallica.bnf.fr/ark:/12148/cb32695639f/date19010101
Use cases: Time Machine (Maps of Space and Time)

Source:
Frederic Kaplan's TEDx Talk (June 2013)
Use cases: Wiegand Bibliography

● Lindemann and Khemakhem 2018
Use cases: Wiegand Bibliography

- Lindemann and Khemakhem 2018
Conclusion & Outlook

● New approach for structuring Lexical, Encyclopedic and Bibliographic information in entry based documents
● Standardised output to enable exchangeability and enrichment

● More feature engineering is required
  ○ Collect more varied data through GROBID-Dictionaries Workshop Series
● New techniques could be also tested for a hybrid processing
References

- **Sean Crist (2011).** Processing the text of bilingual print dictionaries
- **Mohamed Khemakhem, Axel Herold, Laurent Romary (2018a).** Enhancing Usability for Automatically Structuring Digitised Dictionaries. GLOBALEX workshop at LREC 2018, May 2018, Miyazaki, Japan. 2018
- **Mohamed Khemakhem, Laurent Romary, Simon Gabay, Hervé Bohbot, Francesca Frontini, et al.(2018b).** Automatically Encoding Encyclopedic-like Resources in TEI. The annual TEI Conference and Members Meeting, Sep 2018, Tokyo, Japan
- **Mohamed Khemakhem, Carmen Brando, Laurent Romary, Frédérique Mélanie-Becquet, Jean-Luc Pinol (2018c).** Fueling Time Machine: Information Extraction from Retro-Digitised Address Directories. JADH2018 "Leveraging Open Data", Sep 2018, Tokyo, Japan
- **David Lindemann and Mohamed Khemakhem (2018).** Retro-digitizing and Automatically Structuring a Large Bibliography Collection. European Association for Digital Humanities (EADH) Conference, December 2018, Galway, Ireland
Thank you for your attention :)

Questions?