PDFAnno: PDFドキュメントのための言語アノテーションツール

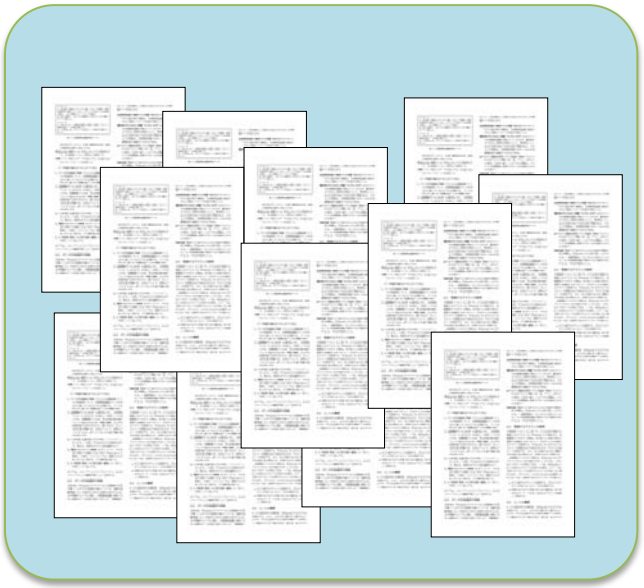
進藤 裕之　　松本 裕治
奈良先端科学技術大学院大学

2017-03-15
言語処理学会第23回年次大会

背景

膨大な科学技術論文からの知識獲得

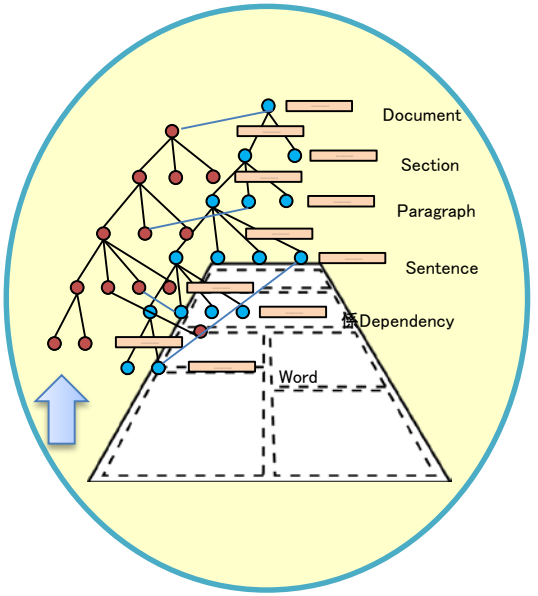
科学技術論文（PDF）



抽出

検索

知識データベース



背景

PDF中のテキストに対するアノテーション

- 目的: 情報抽出・知識獲得のための教師データ作成

- 対象とする主なアノテーション項目:
  - 固有表現，専門用語の範囲
  - エンティティ間の「関係」
  - 共参照関係

背景

PDF中のテキストに対するアノテーション

- 多くの学術論文は，PDFファイルのみが入手可能

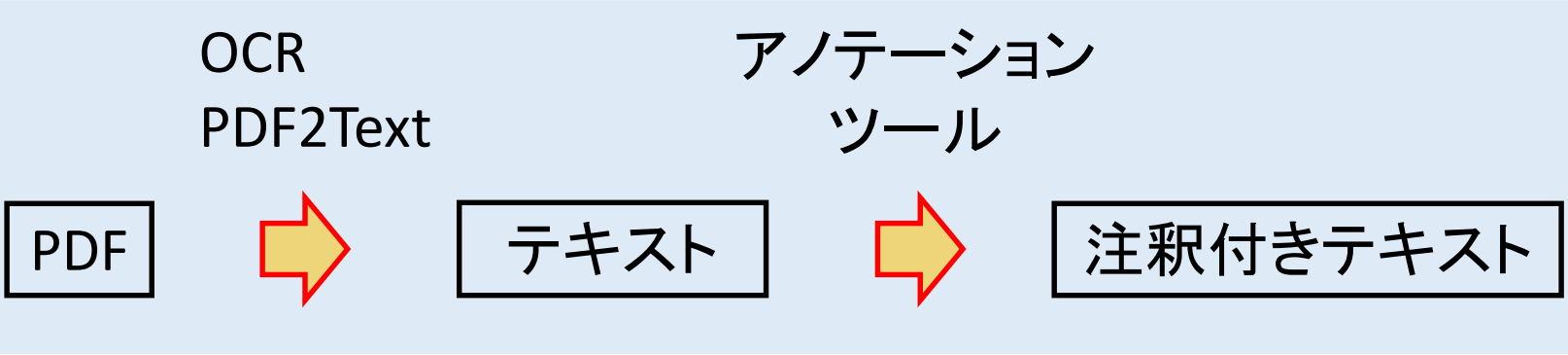
→ PDF中のテキストに対してどのようにアノテーションを行うか？

背景

既存ツールを用いたPDFアノテーション

- 既存の汎用アノテーションツール
  - Brat [Stenetorp et al. 2012]
  - WebAnno [Yimam et al. 2013]
  - MMax2 [Muller et al. 2006]

Plainテキストを対象



※ Adobe Acrobatのような商用PDFソフトウェアは，テキストハイライト，コメント付与は可能だが，関係アノテーション付与は想定されていない

背景

問題点

1. PDF → Textの変換は誤りが多く含まれる
    - 本文と図表・数式が混在してしまう
    - 添え字が上手く認識できない

※ 図（ベクター形式）や表，数式は，PDF中ではテキストとして埋め込まれているので，本文と区別することは簡単でない．

背景

問題点

2. 変換後のPlainテキストは，PDFの段落構造が欠如している
   - アノテーターが文章読解に時間を要するため，アノテーション効率が著しく低下する

問題点

論文PDFの例

Chem. Mater. **2000,** *12,* 2424−2427

Synthesis and Thermoelectric Properties of the New Oxide Materials $Ca_{3-x}Bi_xCo_4O_{9+\delta}$ ($0.0 < x < 0.75$)

Siwen Li,* Ryoji Funahashi, Ichiro Matsubara, Kazuo Ueno, Satoshi Sodeoka, and Hiroyuki Yamada

Department of Energy Conversion, Osaka National Research Institute, AIST, Midorigaoka 1-8-31, Ikeda, Osaka 563, Japan

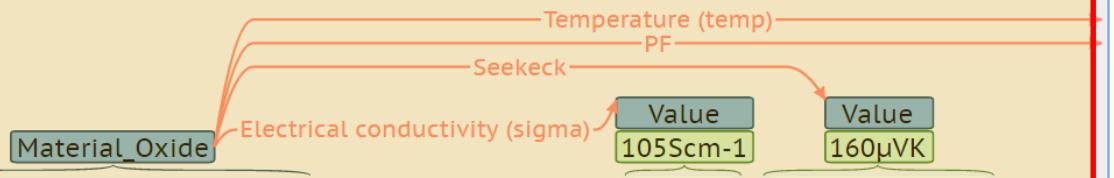
Received February 14, 2000. Revised Manuscript Received May 30, 2000

A new series of oxides $Ca_{3-x}Bi_xCo_4O_{9+\delta}$, ($x = 0.0-0.75$) with $Ca_2Co_2O_5$-type structures were synthesized, and their structures, electrical properties, Seebeck coefficients, and thermal conductivities were measured. The values of Seebeck coefficients of the new oxides are all positive, showing that they are p-type conductors. Both the electrical conductivity and Seebeck coefficients increase with the increasing Bi contents which can be attributed to the increase of carrier mobility due to the larger size of Bi ion. The electrical conductivity, Seebeck coefficient, and the calculated value of the power factor of $Ca_{3-x}Bi_xCo_4O_{9+\delta}$ ($x = 0.5$) are 105 S cm^{-1}, 160 μV K^{-1}, and 2.7×10^{-4} W K^{-2} m^{-1} at 700 °C, respectively. The thermal conductivity of $Ca_{3-x}Bi_xCo_4O_{9+\delta}$ ($x = 0.5$) at room temperature is 1.14 W m^{-1} K^{-1} and increase slightly with the increasing temperature. At 700 °C, the figure of merit of $Ca_{3-x}Bi_xCo_4O_{9+\delta}$ ($x = 0.5$) is 2.0×10^{-4} K^{-1}.

問題点

WebAnnoのアノテーション作業画面

2424
Chem. Mater. 2000, 12, 2424-2427
Synthesis and Thermoelectric P roperties of the New Oxide Materials Ca3-xBixCo4O9+δ(0.0<x<0.75)
Siwen Li, *Ryoji Funahashi, Ichiro Matsubara, Kazuo Ueno, Satoshi Sodeoka, and Hiroyuki Yamada
Department of Energy Conversion, Osaka National Research Institute, AIST, Midorigaoka 1-8-31, Ikeda, Osaka 563, Japan
Received February 14, 2000. Revised Manuscript Received May 30, 2000
Anew series ofoxides CaBiCoO, (x=0.O-O.75) with CaC2o2O5-type structures were synthesized, and their structures, electrical properties, Seebeck coefcients, and thermal conductivities were measured. The values of Seebeck coefcients of the new oxides are all positive, showing that they are p-type conductors. Both the electrical conductivity and Seebeck coefcients increase with the increasing Bi contents which can be attributed to the increase ofcarrier mobility due to the larger size ofBi ion. The electrical conductivity, Seebeck coefficient, and the

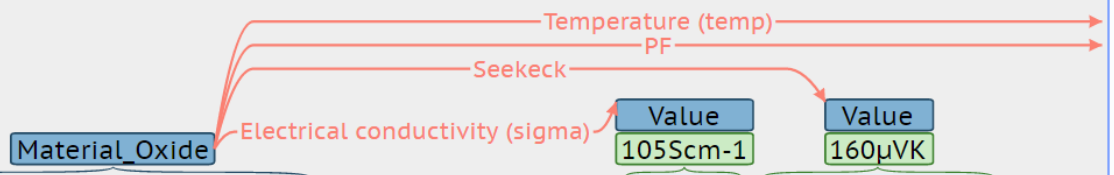


calculated value ofthe power factor of Ca3-xBixCo4O9+δ(x=0.5) are 105 S c m-1,160μVK-1, and

問題点

WebAnnoのアノテーション作業画面

2 2424
3 Chem. Mater. 2000, 12, 2424-2427
4 Synthesis and Thermoelectric P roperties of the New Oxide Materials Ca3-xBixCo4O9+δ(0.0<x<0.75)
5 Siwen Li, *Ryoji Funahashi, Ichiro Matsubara, Kazuo Ueno, Satoshi Sodeoka, and Hiroyuki Yamada
6 Department of Energy Conversion, Osaka National Research Institute, AIST, Midorigaoka 1-8-31, Ikeda, Osaka 563, Japan
7 Received February 14, 2000. Revised Manuscript Received May 30, 2000
8 Anew series ofoxides CaBiCoO, (x=0.O-O.75) with CaC2o2O5-type structures were synthesized, and their structures, electrical properties, Seebeck coefcients, and thermal conductivities were measured. The values of Seebeck coefcients of the new oxides are all positive, showing that they are p-type conductors. Both the electrical conductivity and Seebeck coefcients increase with the increasing Bi contents which can be attributed to the increase ofcarrier mobility due to the larger size ofBi ion. The electrical conductivity, Seebeck coefficient, and the



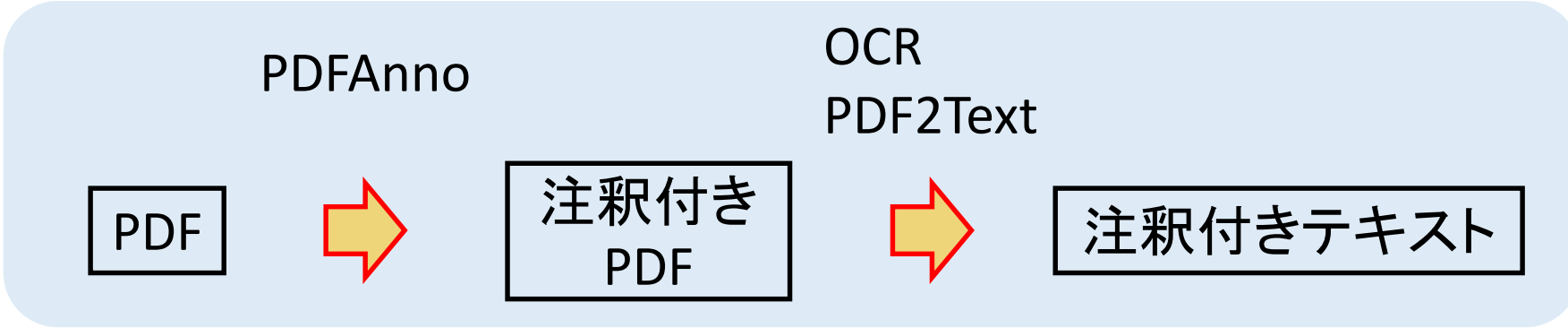
calculated value ofthe power factor of Ca3-xBixCo4O9+δ(x=0.5) are 105 S c m-1,160μVK-1, and

実際には，PDFとWeb画面を見比べながらアノテーション作業を行う必要がある

→ 作業効率の低下，アノテーション誤りの増加

解決策

PDF上に直接アノテーションを行う



解決策

PDF上に直接アノテーションを行う

メリット:

- 既存ツールを用いた場合の問題点を解消できる
- PDF → Text変換ソフトの性能が向上した場合に，アノテーションデータの修正が不要

（関連研究）

1. ACL論文コーパスの構築 [Steven et al. 2008]
2. ACLコーパスに対する共参照アノテーション [Schafer et al. 2012]
    - 1と2では，PDF → Text変換ソフトが異なる

解決策

PDF上に直接アノテーションを行う

デメリット:

- PDFは行間を調節できない

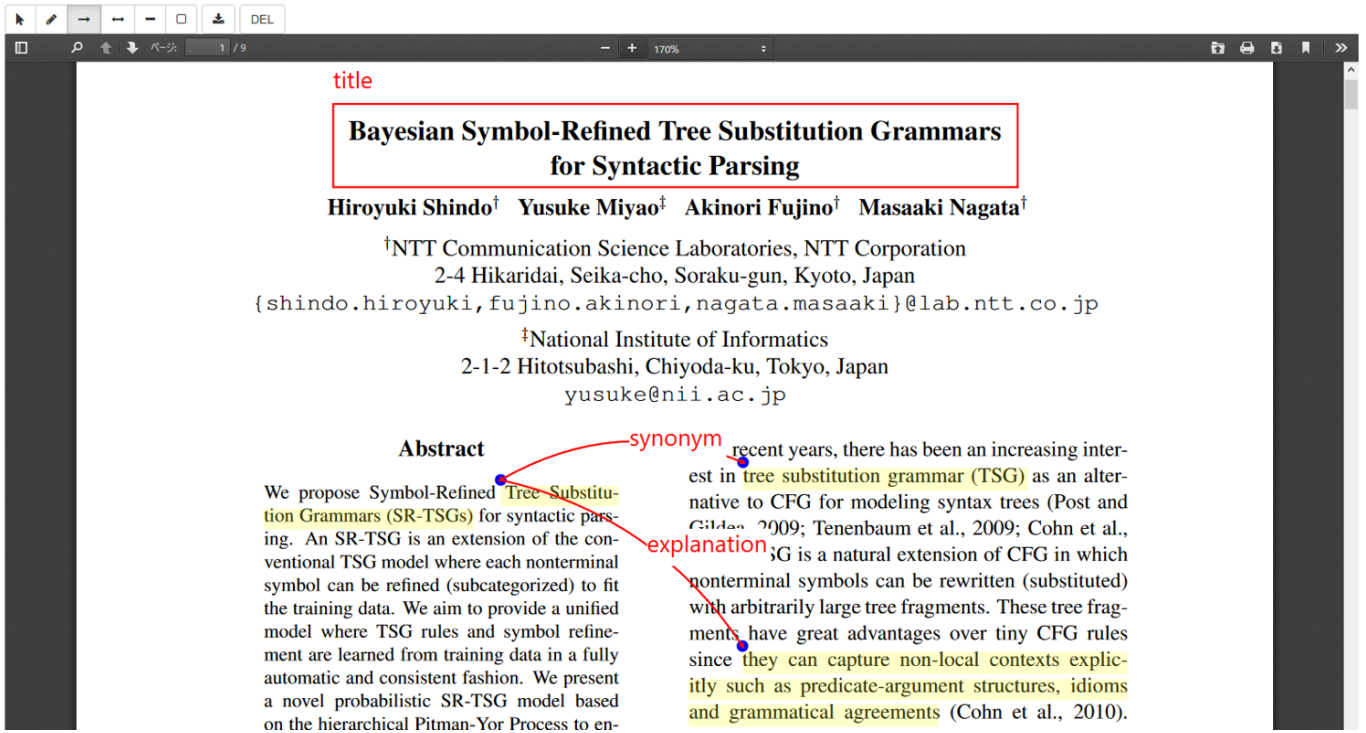
  → 構文木などのアノテーションには不適

固有表現認識・関係抽出などの正解データ作成に最適

PDFAnno

アノテーションの種類

- Span
- Rectangle
- Relation（単方向，双方向，無方向）



PDFAnno

主な仕様

- Webブラウザ上で動作するPDFアノテーションツール
- Client-onlyアプリケーション（オフラインで動作可）
  ⇔　Brat，WebAnnoは，サーバーと常に通信が必要
- PDF.js（developed by Mozilla）でPDFを描画
- アノテーション情報の描画，ユーザーインターフェースを
  JavaScript + HTML5で実装
- オープンソース（Github），MITライセンス

PDFAnno

アノテーションデータの保存

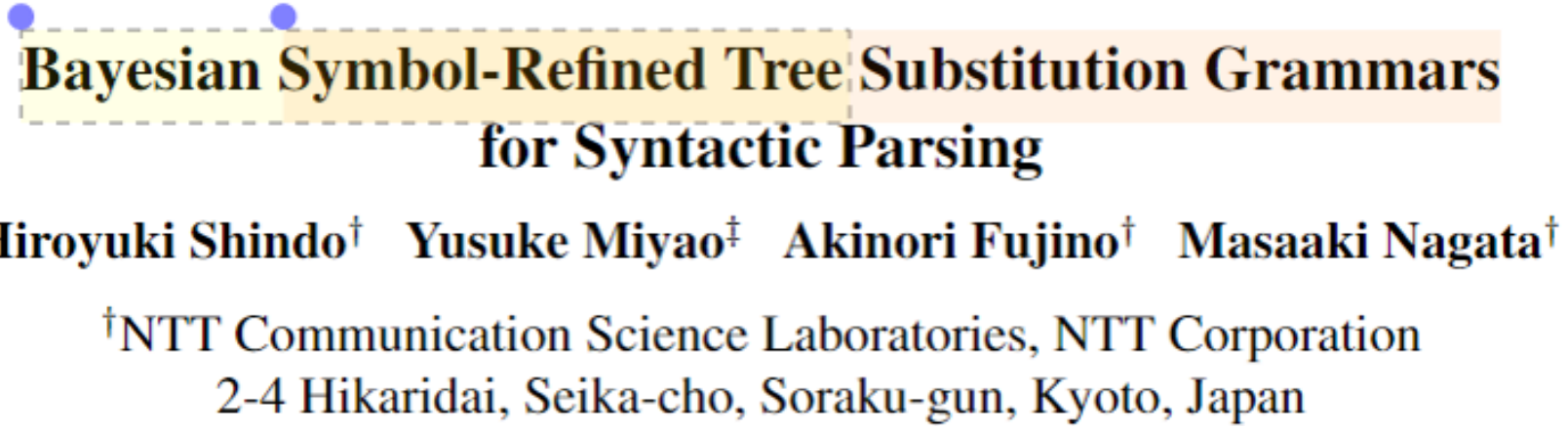
- アノテーションデータは，テキストファイル（TOMLフォーマット）としてダウンロードできる

```
[1]
type = "span"
page = 1
position = [[95.818, 252.977, 181.761, 10.909], [95.818, 264.806, 107.136, 10.909]]
label = "label-1"

[2]
type = "span"
page = 1
position = [[323.863, 230.715, 213.988, 11.590], [313.125, 244.522, 224.829, 10.795]]
label = "label-2"

[3]
type = "relation"
dir = "two-way"
ids = ["1", "2"]
label = "label-4"
```

PDFAnno

Agreementのチェック

- 同じPDFファイルに対する複数のアノテーション結果を異なる色で同時に描画できる
  → Agreementのチェック，修正が可能


Bayesian Symbol-Refined Tree Substitution Grammars for Syntactic Parsing

Hiroyuki Shindo† Yusuke Miyao‡ Akinori Fujino† Masaaki Nagata†

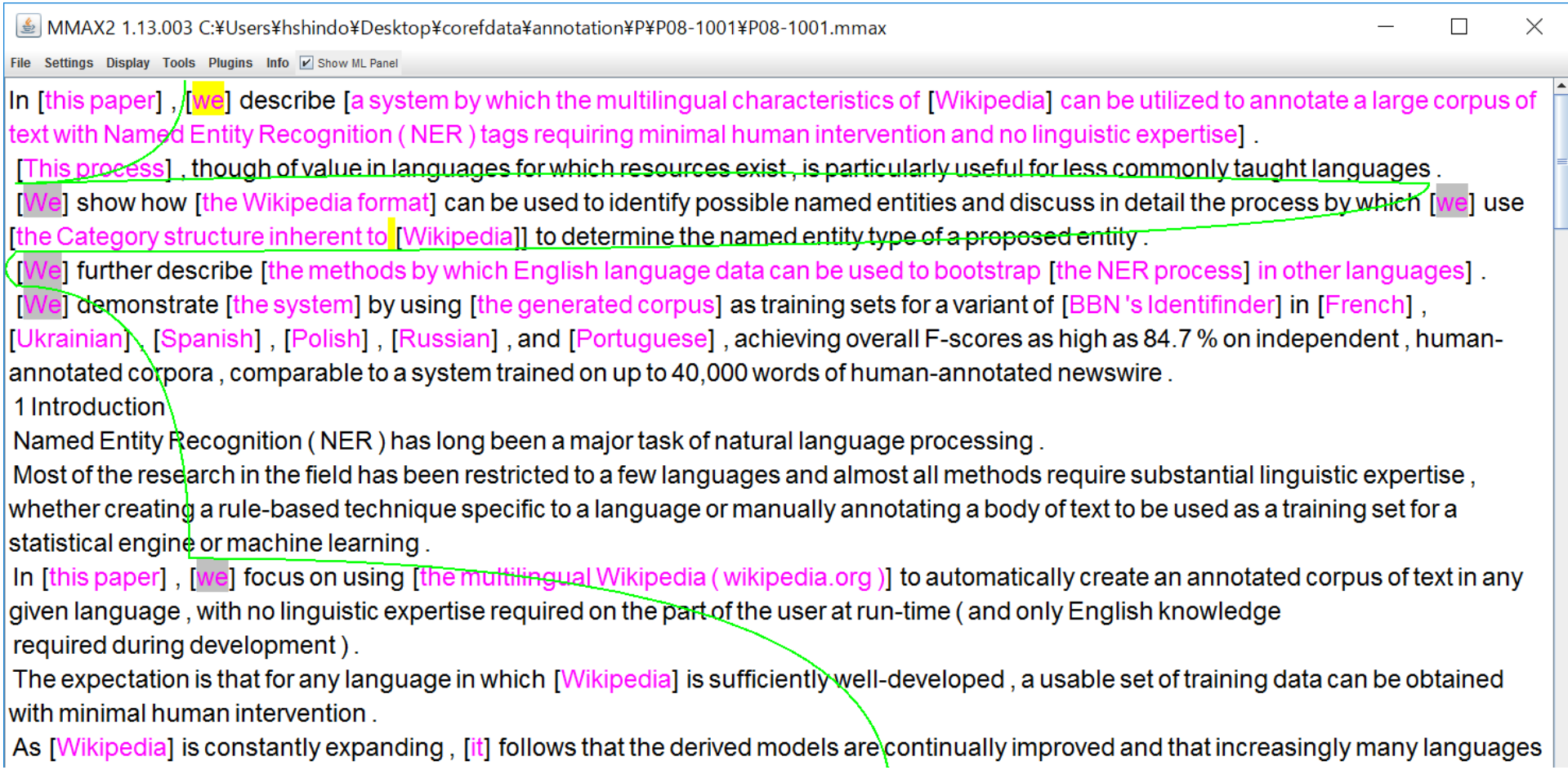
†NTT Communication Science Laboratories, NTT Corporation
2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto, Japan

ケーススタディ

- ACLの文献に対する共参照アノテーション
- 材料工学の文献に対する物質名・物理量のアノテーション

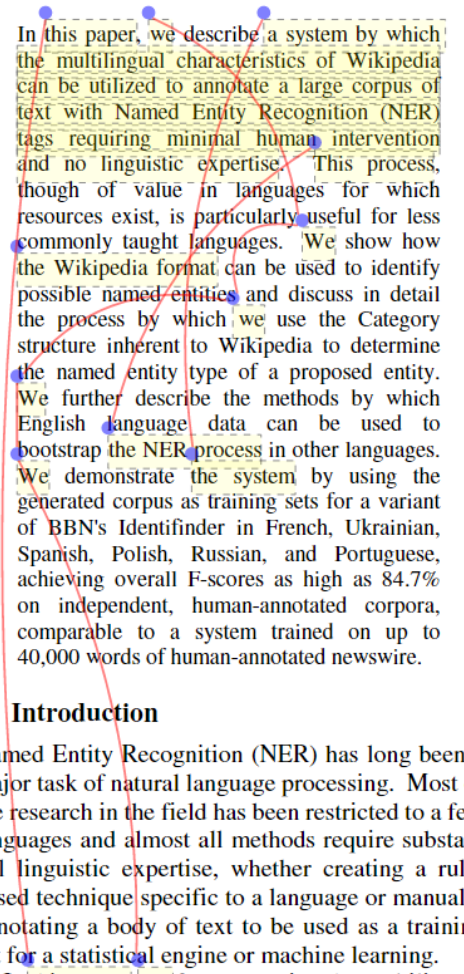
ACL論文の共参照アノテーション

MMax2 [Muller et al. 2006] によるアノテーション作業画面



ACL論文の共参照アノテーション

PDFAnno によるアノテーション作業画面



Abstract

In this paper, we describe a system by which the multilingual characteristics of Wikipedia can be utilized to annotate a large corpus of text with Named Entity Recognition (NER) tags requiring minimal human intervention and no linguistic expertise. This process, though of value in languages for which resources exist, is particularly useful for less commonly taught languages. We show how the Wikipedia format can be used to identify possible named entities and discuss in detail the process by which we use the Category structure inherent to Wikipedia to determine the named entity type of a proposed entity. We further describe the methods by which English language data can be used to bootstrap the NER process in other languages. We demonstrate the system by using the generated corpus as training sets for a variant of BBN's Identifinder in French, Ukrainian, Spanish, Polish, Russian, and Portuguese, achieving overall F-scores as high as 84.7% on independent, human-annotated corpora, comparable to a system trained on up to 40,000 words of human-annotated newswire.

1 Introduction

Named Entity Recognition (NER) has long been a major task of natural language processing. Most of the research in the field has been restricted to a few languages and almost all methods require substantial linguistic expertise, whether creating a rule-based technique specific to a language or manually annotating a body of text to be used as a training set for a statistical engine or machine learning.

In this paper, we focus on using the multilingual Wikipedia (wikipedia.org) to automatically create required during development). The expectation is that for any language in which Wikipedia is sufficiently well-developed, a usable set of training data can be obtained with minimal human intervention. As Wikipedia is constantly expanding, it follows that the derived models are continually improved and that increasingly many languages can be usefully modeled by this method.

In order to make sure that the process is as language-independent as possible, we declined to make use of any non-English linguistic resources outside of the Wikimedia domain (specifically, Wikipedia and the English language Wiktionary (en.wiktionary.org)). In particular, we did not use any semantic resources such as WordNet or part of speech taggers. We used our automatically annotated corpus along with an internally modified variant of BBN's IdentiFinder (Bikel et al., 1999), specifically modified to emphasize fast text processing, called "PhoenixIDF," to create several language models that could be tested outside of the Wikipedia framework. We built on top of an existing system, and left existing lists and tables intact. Depending on language, we evaluated our derived models against human or machine annotated data sets to test the system.

2 Wikipedia

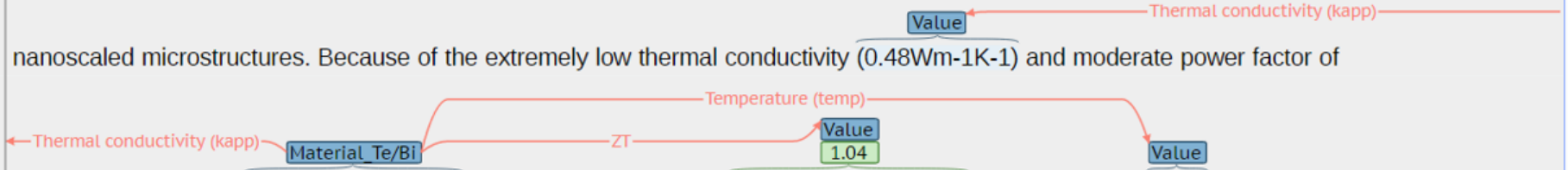
2.1 Structure

Wikipedia is a multilingual, collaborative encyclopedia on the Web which is freely available for research purposes. As of October 2007, there were over 2 million articles in English, with versions available in 250 languages. This includes 30 languages with at least 50,000 articles and another 40 with at least 10,000 articles. Each language is available for download (download.wikimedia.org) in a text format suitable for inclusion in a database

材料工学論文のアノテーション

WebAnno [Yimam et al. 2013] によるアノテーション作業画面

19 1Department of Engineering and System Science, National Tsing Hua University, Hsinchu, Taiwan
20 2Institute of Physics, Academia Sinica, Taipei, Taiwan
21 3Nano Science and Technology, Taiwan International Graduate Program, Academia Sinica, Taipei, Taiwan 4Center for Condensed Matter Sciences, National Taiwan University, Taipei, Taiwan
22 5National Synchrotron Radiation Research Center, Hsinchu, Taiwan
23 6Center for Emerging Material and Advanced Devices, National Taiwan University, Taipei, Taiwan
24 7 Graduate Institute of Applied Physics, National Chengchi University, Taipei, Taiwan
25 (Received 20 May 2013; accepted 16 October 2013; published online 31 October 2013)
26 The influence of bismuth (Bi) substitution on the thermoelectric properties of AgSbTe2 compounds was investigated and compared with the undoped AgSbTe. The addition of Bi dopants not only resulted in a reduction in thermal conductivity but also markedly increased the thermopower in the Ag(Sb1-xBix)Te2 series. Additional phonon scatterings were created by Bi doping and led to a reduction of thermal conductivity. The lattice thermal conductivity is significantly reduced which could be ascribed to enhancement of phonon scattering by dopants with greater atomic weight. In addition, the thermopower was enhanced, which was attributed to the electron-filtering effects caused by the



greater than 10% compared with an undoped AgSbTe2. this result shows promising thermoelectric properties in the medium temperature range. ©2013 AIP Publishing LLC. [http: //dx.doi.org/10.1063/1.4828478]

27 I. INTRODUCTION
28 Energy and the environment have become critical issues in the 21st century. The urgent need for sources of energy other than fossil fuels, as well as the most efficient use of current fossil-fuel supply, has prompted urgent research into alternative energy sources and various types of energy conversion technologies. One type of energy conversion technology that has gained renewed attention is thermoelectric (TE) energy conversion, where heat is converted directly into electricity using a class of materials known as thermoelectric materials.1-7 TE materials are

※ 「0.48Wm-1K-1」のように，添え字が崩れている

材料工学論文のアノテーション

PDFAnno によるアノテーション作業画面

Rajeshkumar Mohanraman,[1,2,3,a)] Raman Sankar,[4] F. C. Chou,[4,5,6] C. H. Lee,[1,5] and Yang-Yuan Chen[2,7,a)]

[1]*Department of Engineering and System Science, National Tsing Hua University, Hsinchu, Taiwan*
[2]*Institute of Physics, Academia Sinica, Taipei, Taiwan*
[3]*Nano Science and Technology, Taiwan International Graduate Program, Academia Sinica, Taipei, Taiwan*
[4]*Center for Condensed Matter Sciences, National Taiwan University, Taipei, Taiwan*
[5]*National Synchrotron Radiation Research Center, Hsinchu, Taiwan*
[6]*Center for Emerging Material and Advanced Devices, National Taiwan University, Taipei, Taiwan*
[7]*Graduate Institute of Applied Physics, National Chengchi University, Taipei, Taiwan*

(Received 20 May 2013; accepted 16 October 2013; published online 31 October 2013)

The influence of bismuth (Bi) substitution on the thermoelectric properties of $AgSbTe_2$ compounds was investigated and compared with the undoped $AgSbTe_2$. The addition of Bi dopants not only resulted in a reduction in thermal conductivity but also markedly increased the thermopower in the $Ag(Sb_{1-x}Bi_x)Te_2$ series. Additional phase scattering centers are created by Bi doping and led to a reduction of thermal conductivity. The lattice thermal conductivity is significantly reduced which could be ascribed to enhancement of phonon scattering by dopants with greater atomic weight. In addition, the thermopower was enhanced, which was ascribed to the electron-filtering effects caused by the nanoscaled microstructures. Because of the extremely low thermal conductivity ($0.48\ Wm^{-1}K^{-1}$) and moderate power factor of $AgBi_{0.05}Sb_{0.95}Te_2$, a maximum ZT value of (1.04 ± 0.08) was reached at 570 K; yielding an enhancement of greater than 10% compared with an undoped $AgSbTe_2$. this result shows promising thermoelectric properties in the medium temperature range. © *2013 AIP Publishing LLC*. [http://dx.doi.org/10.1063/1.4828478]

Temperature (temp)

Material_Te/Bi

Value

まとめ

- PDFAnno: PDF上にテキストアノテーションを行うツール
- PDF上に直接アノテーションすることによって
  - PDF → Text変換の精度に左右されない
  - PDFの段落構造が保存されるため効率的に作業できる
- フリー，オープンソースで公開

  https://github.com/paperai/pdfanno