SYNGRAPH: 国語辞典とコーパスから自動抽出した同義・上位下位関係に基づく柔軟マッチング

柴田 知秀　　小谷 通隆　　黒橋 禎夫

平成 20 年 8 月 7 日

1　概要

[4, 3, 2] を参照。

2　国語辞典の整形・マージ

3　国語辞典からの語の関係の抽出

国語辞典・コーパスから類義表現を抽出し、ゴミの削除、マージ、多義性解消などを行なう。
国語辞典から、以下の関係を抽出する。

- 同義表現
- 上位下位関係
- 反義関係
- 同義句

3.1　上位下位関係

3.2　同義表現

3.3　反義関係

辞書の記号「反対語」を利用し、反義関係を抽出する。以下の例では、「みぎ」の反対語として「左」が抽出される。

みぎ
　北を向いたときに、東にあたるほう。
　<用例>右がわ。
　<反対語>左。</反対語>

3.4　同義句

4　コーパスからの同義表現抽出

4.1　括弧表現を利用したコーパスからの同義表現抽出

笹野らの手法 [1] を利用する。

5　類義表現の整理

以下のコマンドで行なうことができる。

```
% cd scripts
% ./merge_dic.sh
```

オプションを以下に示す。

- -m: 人手での修正を反映
- -w: Wikipedia 用

5.1　同義表現のマージ

複数の語が同一である類義表現のグループを、一つの類義表現のグループにマージする。副産物として、多義性が解消できる (場合がある)。

以下の 2 つの類義表現のグループを、

- 書籍/しょせき:1/1:1/1 本/ほん 書物/しょもつ 図書/としょ
- 書物/しょもつ:1/1:1/1 本/ほん 書籍/しょせき

以下のようにマージする。

- 書物/しょもつ:1/1:1/1 本/ほん 書籍/しょせき:1/1:1/1 図書/としょ

6　同義表現データベースのコンパイル

以下のコマンドで行なうことができる。

```
% cd scripts
% ./build.sh
```

主なオプションを以下に示す。

- -o: orchid
- -l: CGI 用
- -j: JUMAN コマンドを指定
- -r: JUMANRC を指定
- -d: knp を-dpnd で動かす (デフォルトは格解析)
- -w: Wikipedia 用

以下が順に行なわれる。

- 同義グループに SYNID 付与
- 構文解析
- コンパイル

以下では各手順について詳しく述べる。

6.1 同義グループに SYNID 付与

6.2 構文解析

6.3 コンパイル

7 SYNGRAPH フォーマット

KNP の解析結果に SYNGRAPH の情報をうめこむことができる。図 1 に「ホテルに一番近い駅」の解析結果を示す。

以下に notation の説明を示す。

- 「#」、「*」、「+」行は KNP の解析結果と同じ。
- 「!!」が付いている行は同じ基本句に対応している基本ノード、SYN ノードに共通する情報を出力する。 左から順に、対応している基本句番号、親のノードが対応している基本句番号（複数ある場合は「/」でつないでいる）、係り方、見出し、さらに存在すれば文法フラグ、格解析結果など。
- 「!」が付いている行は各ノードの情報を出力する。左から順に、対応する基本句番号、SYNID（基本 ID も SYNID として表記）、スコア、さらに存在すれば文法素性、上位語、反義語などが出力される。以下に文法素性を示す。
  - 否定
  - 可能
  - 尊敬
  - 受身
  - 使役

  その他に、あれば下位語数が出力される。

以下のサンプルプログラムでテストすることができる。

```
% cd scripts
% perl knp_syn.pl -s ホテルに一番近い駅
```

主なオプションを以下に示す。

- -relation: 上位語を付与
- -antonym: 反義語を付与
- -dbdir: 同義表現データベースを指定

図 1 に示した解析結果は以下のコマンドで動かしたものである。

```
% perl knp_syn.pl -s ホテルに一番近い駅 -relation
```

```
# S-ID:1 KNP:3.0-20080214 DATE:2008/08/06 SCORE:-30.72806 SynGraph:1.7-20080805
* 2D <SM-主体><SM-場所><SM-組織><BGH:ホテル/ほてる><文頭><ニ><助詞><体言><係:ニ格><区切:0-0><RID:1180><格要素><連用要素><正規化代表表記:ホテル/ほてる><主辞代表表記:ホテル/ほてる>
+ 2D <SM-主体><SM-場所><SM-組織><BGH:ホテル/ほてる><文頭><ニ><助詞><体言><係:ニ格><区切:0-0><RID:1180><格要素><連用要素><名詞項候補><先行詞候補><正規化代表表記:ホテル/ほてる><解析格:ニ>
ホテル ほてる ホテル 名詞 6 普通名詞 1 * 0 * 0 "組織名末尾 カテゴリ:場所-施設 ドメイン:レクリエーション:ビジネス 代表表記:ホテル/ほてる" <組織名末尾><カテゴリ:場所-施設><ドメイン:レクリエーション:ビジネス><代表表記:ホテル/ほてる><正規化代表表記:ホテル/ほてる><文頭><記英数カ><カタカナ><名詞相当語><自立><内容語><タグ単位始><文節始><固有キー><文節主辞>
に に に 助詞 9 格助詞 1 * 0 * 0 NIL <品曖><ALT-に-に-に-9-3-0-0-NIL><品曖-格助詞><品曖-接続助詞><かな漢字><ひらがな><付属>
!! 0 1,2/2D <見出し:ホテルに><格解析結果:ニ格>
! 0 <SYNID:ホテル/ほてる><スコア:1>
! 0 <SYNID:s9192:宿泊施設><スコア:0.99>
! 0 <SYNID:s1866:ホテル/ほてる><スコア:0.99>
! 0 <SYNID:s249:宿/やど><スコア:0.693><上位語><下位語数:1>
* 2D <BGH:一番/いちばん><相対名詞修飾><用言弱修飾><副詞><修飾><係:連用><区切:0-4><RID:1398><連用要素><正規化代表表記:一番/いちばん><主辞代表表記:一番/いちばん>
+ 2D <BGH:一番/いちばん><相対名詞修飾><用言弱修飾><副詞><修飾><係:連用><区切:0-4><RID:1398><連用要素><正規化代表表記:一番/いちばん><解析格:修飾>
一番 いちばん 一番 副詞 8 * 0 * 0 * 0 "相対名詞修飾 用言弱修飾 代表表記:一番/いちばん" <相対名詞修飾><用言弱修飾><代表表記:一番/いちばん><正規化代表表記:一番/いちばん><漢字><かな漢字><自立><内容語><タグ単位始><文節始><文節主辞>
!! 1 2D <見出し:一番><格解析結果:修飾格>
! 1 <SYNID:一番/いちばん><スコア:1>
! 1 <SYNID:s523:トップ/とっぷ><スコア:0.99>
! 1 <SYNID:s2196:何より/なにより><スコア:0.99>
! 1 <SYNID:s1987:一番/いちばん><スコア:0.99>
! 1 <SYNID:s1988:最も/もっとも><スコア:0.99>
* 3D <BGH:近い/ちかい><連体修飾><用言:形><係:連格><レベル:B-><区切:0-5><ID:（形判連体）><RID:765><連体並列条件><正規化代表表記:近い/ちかい><主辞代表表記:近い/ちかい>
+ 3D <BGH:近い/ちかい><連体修飾><用言:形><係:連格><レベル:B-><区切:0-5><ID:（形判連体）><RID:765><連体並列条件><正規化代表表記:近い/ちかい><用言代表表記:近い/ちかい><格要素-ガ:駅><格要素-ニ:ホテル><格要素-ト:NIL><格要素-デ:NIL><格要素-カラ:NIL><格要素-ヨリ:NIL><格要素-マデ:NIL><格要素-ヘ:NIL><格要素-時間:NIL><格要素-外の関係:NIL><格要素-ノ:NIL><格要素-修飾:一番><格要素-トスル:NIL><格要素-ニクラベル:NIL><格要素-ガ２:NIL><格要素-ヲノゾク:NIL><格フレーム-ガ-主体><格フレーム-ニ-主体><格フレーム-ヨリ-主体><格フレーム-ノ-主体><格フレーム-ニクラベル-主体><格フレーム-ガ２-主体><格フレーム-ガ-主体ｏｒ主体準><格フレーム-ガ２-主体ｏｒ主体準><格フレーム-ニ-主体ｏｒ主体準><格関係0:ニ:ホテル><格関係1:修飾:一番><格関係3:ガ:駅><格解析結果:近い/ちかい:形 6:ガ/N/駅/3/0/?; ニ/C/ホテル/0/0/?; ト/U/-/-/-/-; デ/U/-/-/-/-; カラ/U/-/-/-/-; ヨリ/U/-/-/-/-; マデ/U/-/-/-/-; ヘ/U/-/-/-/-; 時間/U/-/-/-/-; 外の関係/U/-/-/-/-; ノ/U/-/-/-/-; 修飾/C/一番/1/0/?; トスル/U/-/-/-/-; ニクラベル/U/-/-/-/-; ガ２/U/-/-/-/-; ヲノゾク/U/-/-/-/->
近い ちかい 近い 形容詞 3 * 0 イ形容詞アウオ段 18 基本形 2 "代表表記:近い/ちかい" <代表表記:近い/ちかい><正規化代表表記:近い/ちかい><かな漢字><活用語><自立><内容語><タグ単位始><文節始><文節主辞>
!! 2 3D <見出し:近い>
! 2 <SYNID:近い/ちかい><スコア:1>
! 2 <SYNID:s199:親しい/したしい><スコア:0.99>
! 2 <SYNID:s11:付近/ふきん><スコア:0.99>
! 2 <SYNID:s1201:所在/しょざい><スコア:0.693><上位語><下位語数:323>
! 2 <SYNID:s419:近い/ちかい><スコア:0.99>
!! 1,2 3D <見出し:近い>
! 1,2 <SYNID:s21291:最寄り/もより><スコア:0.99>
! 1,2 <SYNID:s1201:所在/しょざい><スコア:0.693><上位語><下位語数:323>
* -1D <SM-主体><SM-場所><SM-組織><BGH:駅/えき><文末><体言><用言:判><体言止><一文字漢字><レベル:C><区切:5-5><ID:（文末）><裸名詞><RID:1470><提題受:30><主節><定義文主辞><正規化代表表記:駅/えき><主辞代表表記:駅/えき>
+ -1D <SM-主体><SM-場所><SM-組織><BGH:駅/えき><文末><体言><用言:判><体言止><一文字漢字><レベル:C><区切:5-5><ID:（文末）><裸名詞><RID:1470><提題受:30><主節><定義文主辞><判定詞><名詞項候補><先行詞候補><正規化代表表記:駅/えき><用言代表表記:駅/えき><格要素-ガ:NIL><格要素-ヲ:NIL><格要素-ニ:NIL><格要素-ト:NIL><格要素-デ:NIL><格要素-カラ:NIL><格要素-ヨリ:NIL><格要素-マデ:NIL><格要素-ヘ:NIL><格要素-時間:NIL><格要素-外の関係:NIL><格要素-ノ:NIL><格要素-修飾:NIL><格要素-ガ２:NIL><格要素-トスル:NIL><格要素-ニトル:NIL><格要素-ニトモナウ:NIL><格要素-ニアワセル:NIL><格フレーム-ガ-主体><格フレーム-ヲ-主体><格フレーム-ニ-主体><格フレーム-デ-主体><格フレーム-カラ-主体><格フレーム-ヨリ-主体><格フレーム-マデ-主体><格フレーム-ノ-主体><格フレーム-修飾-主体><格フレーム-ガ２-主体><格フレーム-ニトル-主体><格フレーム-ガ-主体ｏｒ主体準><格フレーム-ガ２-主体ｏｒ主体準><格フレーム-ヲ-主体ｏｒ主体準><格フレーム-ニ-主体ｏｒ主体準><解析連格:ガ><格解析結果:駅/えき:判 0:ガ/U/-/-/-/-; ヲ/U/-/-/-/-; ニ/U/-/-/-/-; ト/U/-/-/-/-; デ/U/-/-/-/-; カラ/U/-/-/-/-; ヨリ/U/-/-/-/-; マデ/U/-/-/-/-; ヘ/U/-/-/-/-; 時間/U/-/-/-/-; 外の関係/U/-/-/-/-; ノ/U/-/-/-/-; 修飾/U/-/-/-/-; ガ２/U/-/-/-/-; トスル/U/-/-/-/-; ニトル/U/-/-/-/-; ニトモナウ/U/-/-/-/-; ニアワセル/U/-/-/-/->
駅 えき 駅 名詞 6 普通名詞 1 * 0 * 0 "漢字読み:音 地名末尾 カテゴリ:場所-施設 ドメイン:交通 代表表記:駅/えき" <漢字読み:音><地名末尾><カテゴリ:場所-施設><ドメイン:交通><代表表記:駅/えき><正規化代表表記:駅/えき><文末><表現文末><漢字><かな漢字><名詞相当語><自立><内容語><タグ単位始><文節始><文節主辞>
!! 3 -1D <見出し:駅>
! 3 <SYNID:駅/えき><スコア:1>
! 3 <SYNID:s2245:停車場/ていしゃば><スコア:0.99>
EOS
```

図 1: 「ホテルに一番近い駅」の解析結果

参考文献

[1] Ryohei Sasano, Daisuke Kawahara, and Sadao Kurohashi. Improving coreference resolution using bridging reference resolution and automatically acquired synonyms. In *Anaphora: Analysis, Alogorithms and Applications, 6th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC2007)*, 2007.

[2] Tomohide Shibata, Michitaka Odani, Jun Harashima, Takashi Oonishi, and Sadao Kurohashi. SYNGRAPH: A flexible matching method based on synonymous expression extraction from an ordinary dictionary and a web corpus. In *Proceedings of IJCNLP2008*, 2008.

[3] 小谷通隆, 中澤敏明, 柴田知秀, 黒橋禎夫. SYNGRAPH データ構造における述語項構造の柔軟マッチング. 言語処理学会 第 13 回年次大会, pp. 43–46, 3 2007.

[4] 大西貴士, 黒橋禎夫. 国語辞典からの類義表現抽出と SYNGRAPH データ構造による柔軟マッチング. 言語処理学会 第 12 回年次大会, 3 2006.