文献紹介ゼミ

林 秀治

紹介する文献

- Corpus-Based Syntactic Error Detection Using Syntactic Patterns
- Koldo Gojenola, Maite Oronoz

はじめに

- 構文誤りの検出及び訂正の問題は、自然言語処理の初頭から考えられてきた
- 記述や文法など多くの種類の誤りに対して異なる手法が提案
- しかし、ワープロにおける文法チェックはほとんどされていない

文法チェックがされなかった理由

- 不完全なcoverage
  - 構文解析器のいくつかは実文章の一部は解析できるが、構文的な誤りはほぼ無限であると考えられる

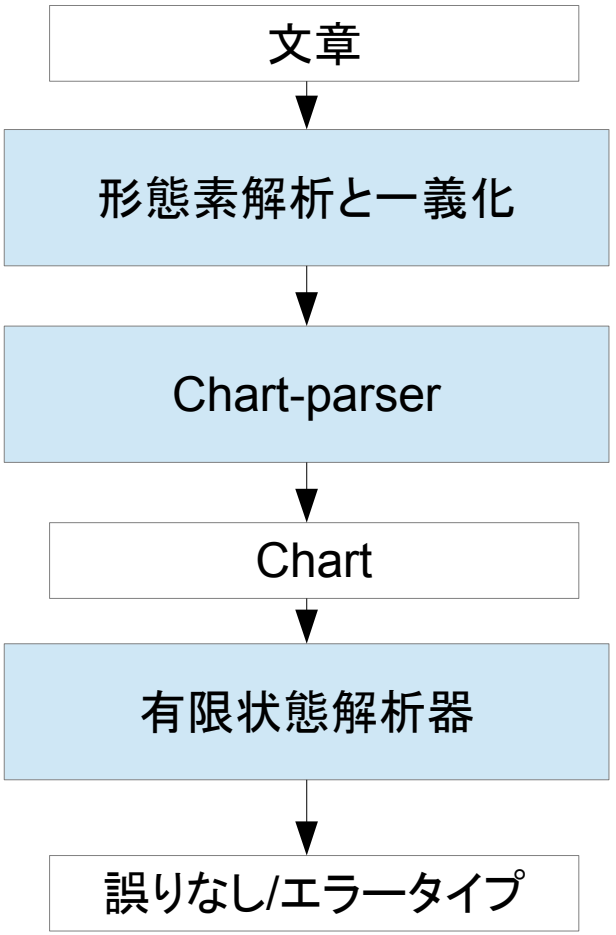
    文が解析できないとき、構文誤りなのか、単にカバーしていないのか判断できない
  - しかし、無制限のコーパスの使用は、正しい構文も誤りとする問題を発生させてしまう

文法チェックがされなかった理由

- 巨大なコーパスの必要性
  - 構文誤りは非常に低頻度で発生する。したがって巨大なコーパスが必要になる。
  - しかし、そのようなコーパスが利用できたとしても誤りを確認する作業は手作業になり、非常にコストがかかる

システムの概要

- 彼らは3つの解析器を使用した



形態素解析と一義化

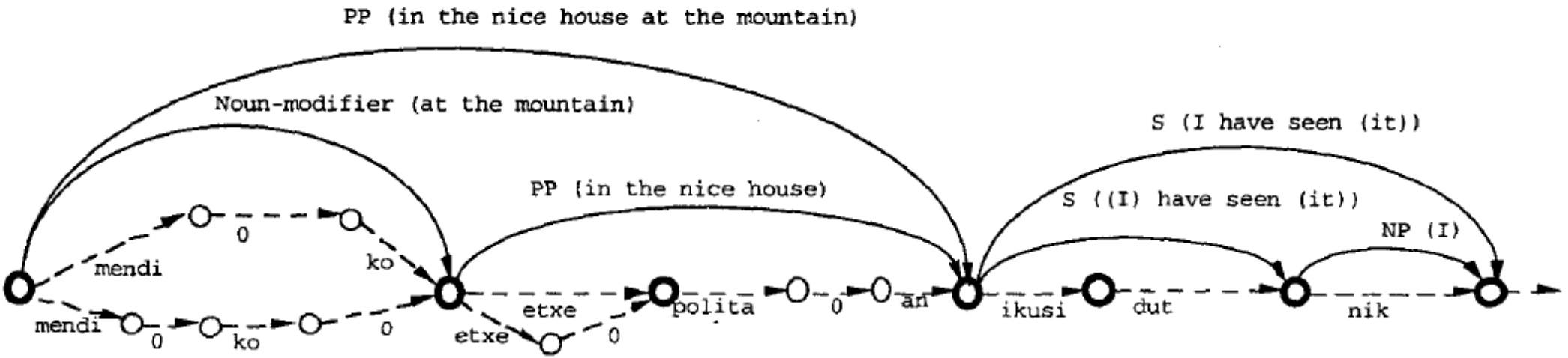
- 形態素解析器によって形態素の構成要素の語ごとの分割を得る
- その後、形態素の一義化が適用され、高レベルの語の曖昧性が解消される

単一化ベースのチャート分析

- 形態素解析と一義化の後、単一化文法はボトムアップで各々の文に適用され、その結果チャートが与えられる

  点線は語彙的な要素、実線は構文の構成要素

  太線の丸は語の境界、通常の丸は形態素の境界を示す



論文より引用

有限状態解析

- チャートは複雑な言語エラーパターンを許容するツールが必要である
- そのため、チャートを有限状態制約を適用できる自動変換器とみることができる

誤り検出

- ほかの誤りに比べ、得るのが比較的簡単なため日付表現を実験対象として選んだ
  - データの大部分は手動で得なければならいが、日付表現は半自動的に獲得するためのいくつかの手掛かり(月の名前や年号など)がある(実験対象はバスク語)
- 人手での評価は検索された文に対して必要

日付表現

- バスク語において大部分の要素は活用しなければならず、対応する数と形態素をつける
- さらに日付の表示形式は関係した要素が固定の組み合わせで現れる
- これらは一般的な誤りのもとで、スペリングチェッカーでは見つけられない

日付表現の例:

Durangon,	1999ko	martxoaren	7an
In Durango,	1999,	March	the 7th

テストデータ

- 学生が書いた226の小論と、新聞・雑誌を集めて50万語以上を得た
- また、そこから、正しい日付や誤った日付、日付に類似したものを含む658文を選んだ
- 人手での選別の結果、誤りの割合は通常のテキストよりも高い
- データは2つにわけ、学習とテストに使われた

エラータイプ

- 誤りの異なる例を調査し、最も頻度の多い6つのエラータイプを調査した

エラータイプ	例
年号はハイフンを使って活用することができない	Donostian, 1995-eko martxoaren 14an
月(martxoak)は小文字で書かれなければならない	1997ko martxoak 14
日付の前の処格(Frantzia)のあとにカンマがなければならない	Frantzia 1997ko irailaren 8an
所有格の月の後の日付(martxoaren)はケースマークを持たなければならい	Donostian, 19995eko martxoaren 22
独立格の月の後の日付(ekainak)はケースマークを持てない	1998.eko ekainak 14ean argitaratua
月(martxoan)は所有格や独立格では活用しなければならない	Donostian, 1995.eko martxoan 28an
エラーの組み合わせ	karrera bukatu nuenean 1997ko Ekainaren 30an

エラータイプ

- 誤りの異なる例を調査し、最も頻度の多い6つのエラータイプを調査した

エラータイプ	例
年号はハイフンを使って活用することができない	Donostian, 1995-eko martxoaren 14an
月(martxoak)は小文字で書かれなければならない	1997ko martxoak 14
日付の前の処格(Frantzia)のあとにカンマがなければならない	Frantzia 1997ko irailaren 8an
所有格の月の後の日付(martxoaren)はケースマークを持たなければならい	Donostian, 19995eko martxoaren 22
独立格の月の後の日付(ekainak)はケースマークを持てない	1998.eko ekainak 14ean argitaratua
月(martxoan)は所有格や独立格では活用しなければならない	Donostian, 1995.eko martxoan 28an
エラーの組み合わせ	karrera bukatu nuenean 1997ko Ekainaren 30an

ルールの改善

- 見られた誤りをもとにエラーに対するルールを考案する
- 正しい日付に対して試験をし、誤検出があればそれがなくなるようにルールを追加

評価

- 開発したコーパスで精度100%、再現率91%を達成できた
- また、初見の247文のコーパスに対しても84%の再現率を得た

	学習に使ったコーパス		テストコーパス	
文章の数	411		247	
検出できなかった誤り	7	9%	6	16%
検出した誤り	84	91%	31	84%
誤検出	0		5	

評価

- テストコーパスで誤検出が5件あったが、正しい日付や日付に類似したものだった
- これをテストコーパス(247文)の数で割れば2.02%ほどであり、結果は有望であるが、精度を向上するためにはより多くのコーパスデータが必要である。