文献紹介ゼミ

林 秀治

紹介する文献

- Plagiarism Detection across Distant Language Pairs
- Alberto Barrón-Cedeňo, Paolo Rosso, Eneko Agirre, Gorka Labaka
- Coling 2010. p37-45

概要

- 単一言語での盗用の検出は様々方法が開発されているのに対し、言語をまたいでの検出は注目されていない
- 2言語間でのT+MA(Translation + Monolingual Analysis),CL-ASA(CL-Alignment-based Similarity Analysis),CL-CNG(CL-Character n-Gram Analysis)の3つのアプローチの有効性を確認する

はじめに

- 盗用(cut-and-paste、挿入、削除及び単語の置換など)は多くの分野で問題になっている
- 様々なアプローチの盗用検出モデルが提案されているが、すべての文書が同じ言語で書かれていることが前提である
- 翻訳過程で生じる盗用もあり、これはcross-language plagiarism(CLP)として知られている

目的

- 現在のCLPの一般的なケースはWeb上の情報で、他言語で書かれた情報を母国語に翻訳し発生することが多い
- 今回は文レベルの盗用を3つのアプローチで検出する

モデルの定義

- T+MA,CL-ASA,CL-CNGの3つの言語間類似度モデルを使用する
- 言語Lで書かれた盗用の疑いのある文dqと、その文のもととなったと思われる言語L'の文書D'を考える

•T+MA

- dqはGiza++,Moses,SRILMによって言語L'のdq'に翻訳される
- 翻訳語はdq'とD'は単一言語として比較可能となる
- そこで単一言語のテキストでの類似性推定に置いて良好な結果が得られたbag-of-wordsを用いたアプローチを使用する
- dq'及びd'の単語の類似度はtf-idfで重み付けされ、コサイン類似度を使って推定する

CL-ASA

- d'はdqの訳である、というような推定をする
- ベイズの定理による機械翻訳を基としている

$$p(d' \mid d_q) = \frac{p(d')\ p(d_q \mid d')}{p(d_q)}.$$

CL-ASA

- 文法的に許容できる翻訳を得るためには、目標言語L'での記述が必要
- dqからLへの変換というよりは、dqの訳をL'で書かれているテキストから検索である
- そこで、言語モデルを長さモデルとして知られるものに取り換える
  - このモデルは言語構造の代わりにテキストの文字長に依存する

$$\varrho(d') = e^{-0.5\left(\frac{\frac{|d'|}{|d_q|}-\mu}{\sigma}\right)^2}$$

μとσはLからL'への翻訳間の文字長の標準偏差

CL-ASA

- 翻訳モデルの確率は以下のように定義される

$$p(d \mid d') = \prod_{x \in d} \sum_{y \in d'} p(x, y)$$

  p(x,y)はxがyの有効な変換である可能性を表す

- CL-ASAに基づく類似度推定は、最終的に以下のように計算される

$$\varphi(d_q, d') = \varrho(d')\ p(d_q \mid d')$$

CL-CGA

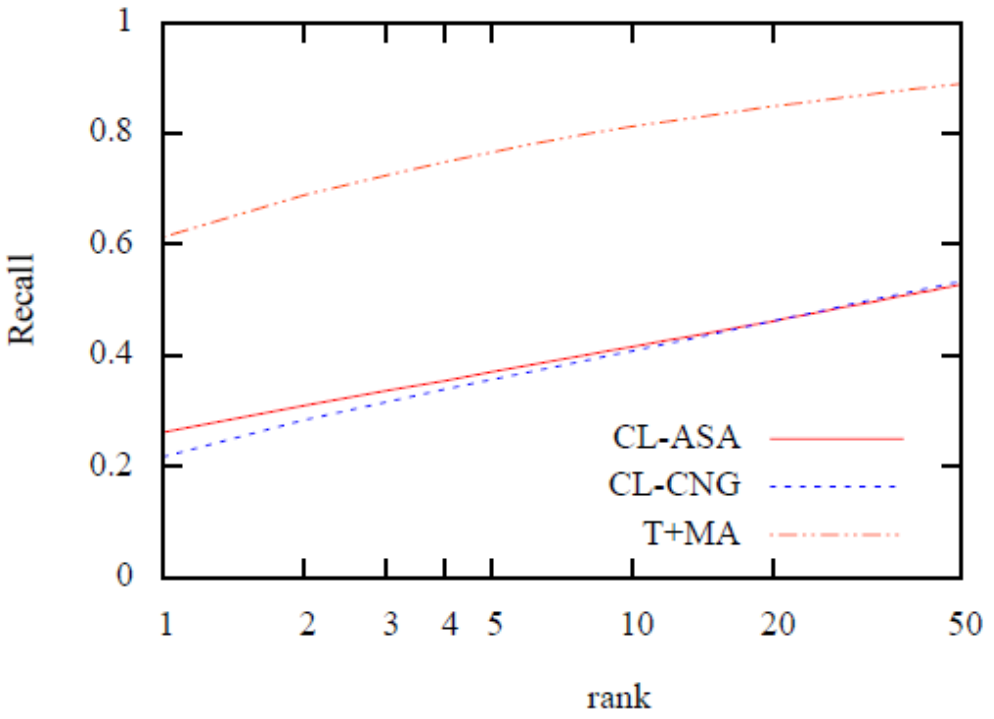
- アルファベットを単純化しΣ={a,…,z,0,…,9}と考える
- ほかの文字は除外する
- 結果として生じたテキストの文字3-gramの出現頻度を用い、tf-idfで重み付けする
- dqとd'の類似度はコサイン類似度によって推定される

実験

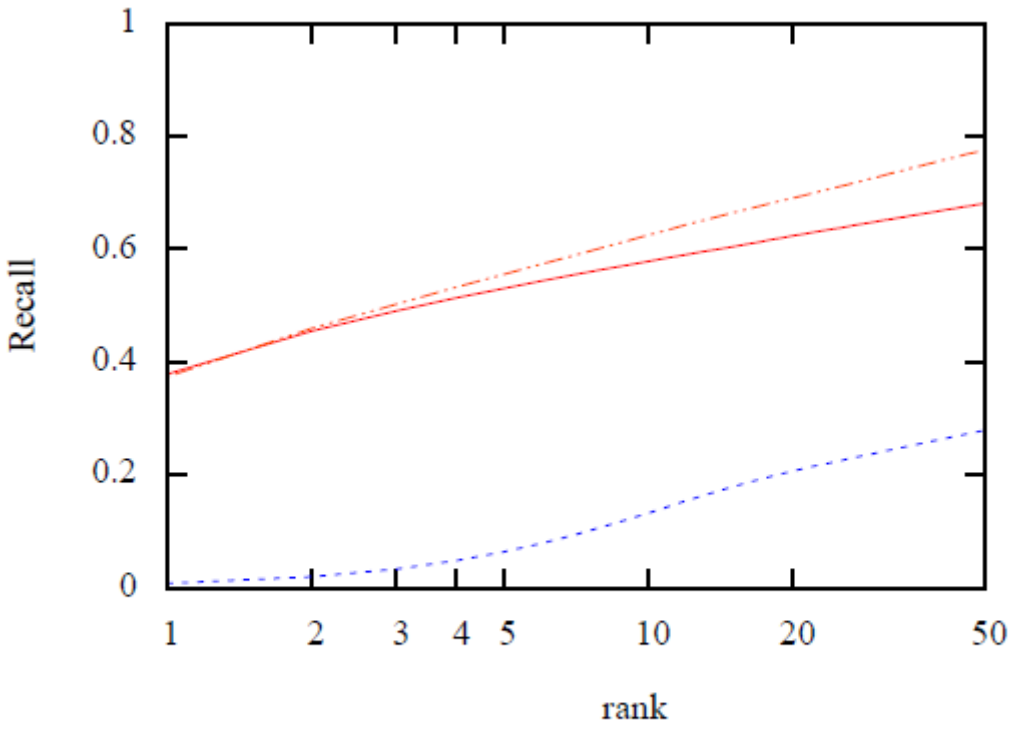
- 3種類のモデルの比較を行った
- 2つのパラレルコーパスを使用
  - Software
    - ソフトフェアマニュアルのen-euの翻訳記録
    - 288,000文
  - Consumer
    - 消費者向け雑誌から抽出されたカタルーニャ語、ガリシア語、バスク語のコーパス
    - 58,202文

結果と考察

- ランク位置n(1~50)の再現率の平均を見た
- 2言語間の構文的な関係が弱いためCL-CNGの結果は悪かった



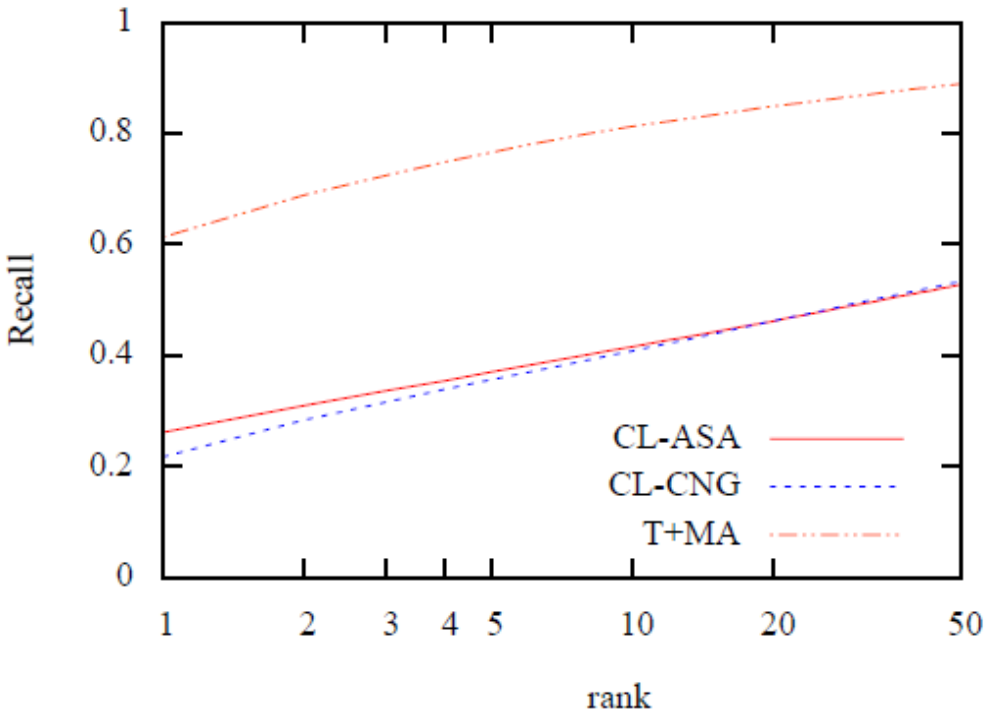
(a) es-eu



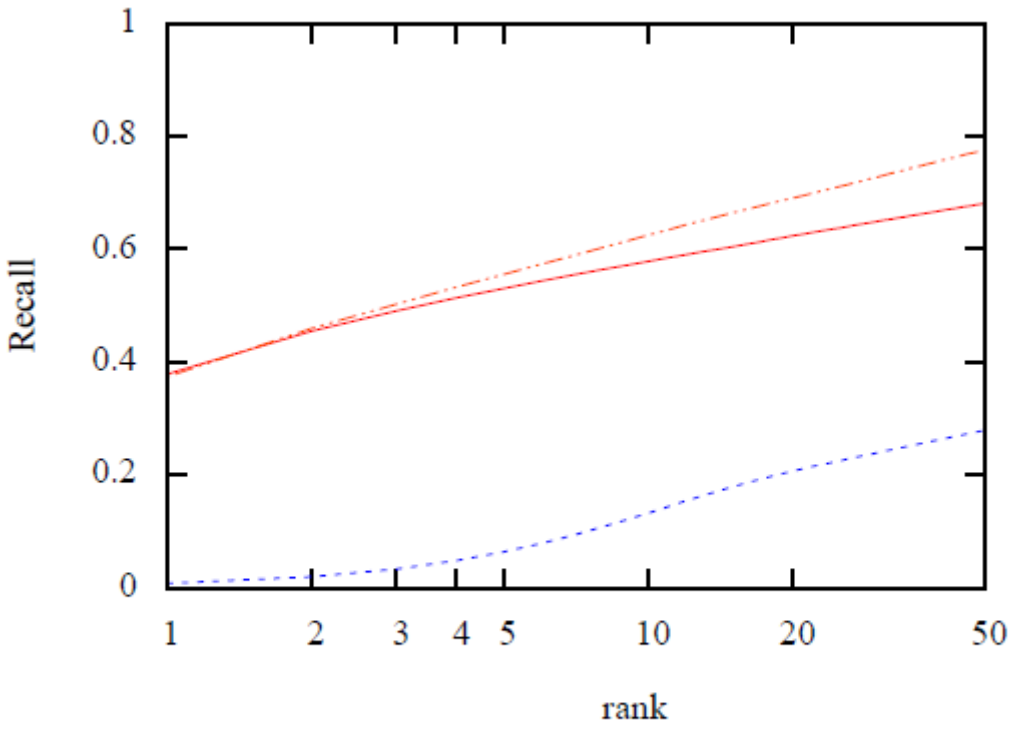
(b) en-eu

結果と考察

- CL-ASAは言語で大きく結果が異なるが、これは利用できるコーパスサイズによるもの(eneuはeseuの約5倍)



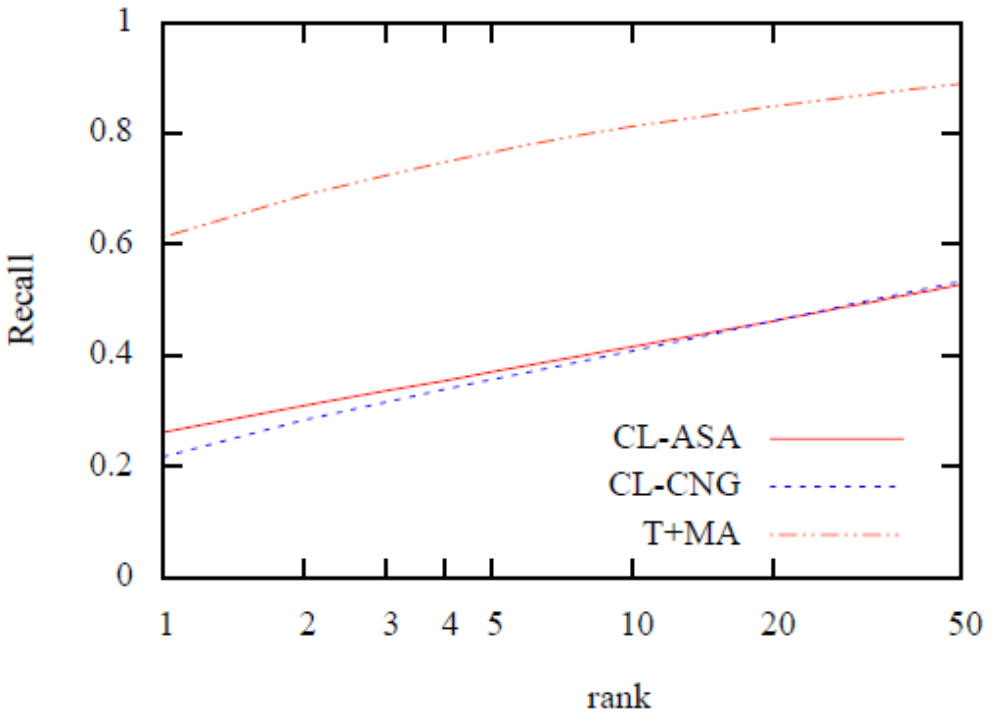
(a) es-eu



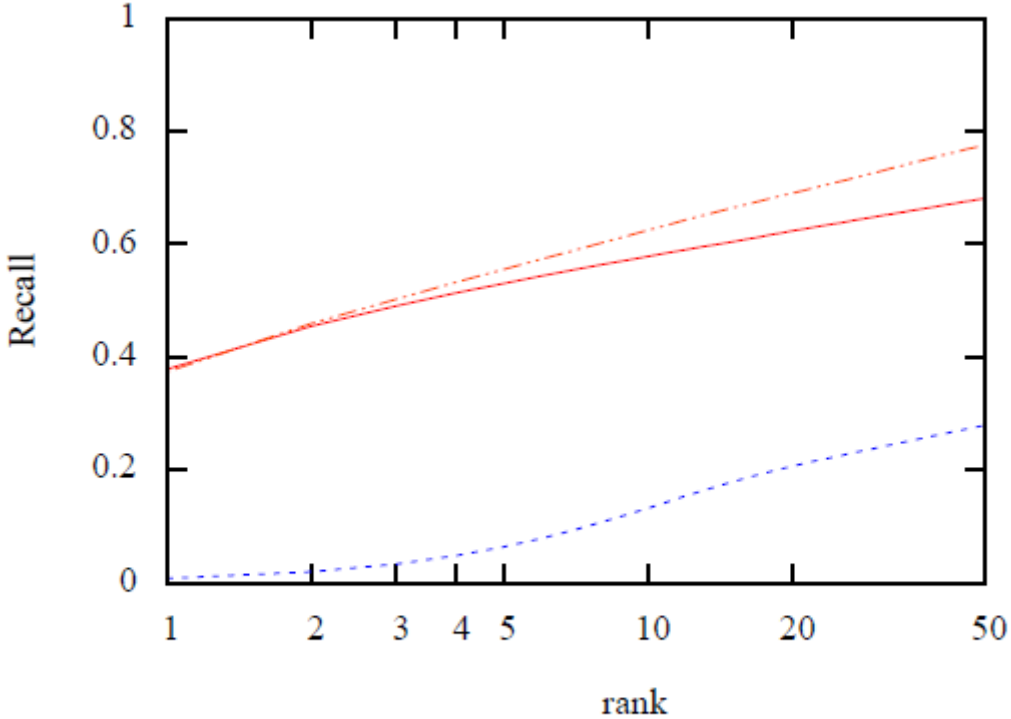
(b) en-eu

結果と考察

- T+MAは翻訳モデルの両方の方向を考慮することで良い結果が得られた



(a) es-eu



(b) en-eu

結果と考察

- T+MAは翻訳モデルの両方の方向を考慮することで良い結果が得られた
- しかしT+MAはすべてのテキストの翻訳を必要とし、非常にコストがかかる
- CL-CNGは、文字列の比較を計算するだけなので非常に用意
- CL-ASAも高速に計算が可能だが、より結果を得るためには大量の計算が必要になる
- 今後はこのトレードオフの検討