背景と目的
まとめ

Wikipedia を用いたソーシャルメディアからの言語横断的な話題抽出システムの試作

中村達哉　白川真澄　原隆浩　西尾章治郎

蒲田　涼馬 (Ryoma Gamada)
u020010@st.pu-toyama.ac.jp

富山県立大学 工学部 情報システム工学科 4 年

April 14, 2023

背景と目的

背景と目的
まとめ

背景

近年では文書集合に含まれるトピックを抽出する研究は数多く行われている. Twitter などのソーシャルメディアがその対象として注目を集めているが, ユーザが自身の言語で情報発信をする多言語なメディアであるため様々な問題が発生する.

目的

多言語なソーシャルメディアを対象として, ソーシャルメディア上で多くの人に言及され話題となっているトピック情報を言語的に抽出・可視化することを目的としたシステムを試作することを目的とする.

試作システムの概要・流れ

背景と目的
まとめ

試作システムの概要

任意の英語の Wikipedia の記事をクエリとして, その記事を中心として Twitter 上で話題となっているトピックを表示するというもの. これによってユーザは関心のあるトピックについての情報を入手し, それについての言語間での共通性や差異を効率的に調べることができる.

試作システムの流れ

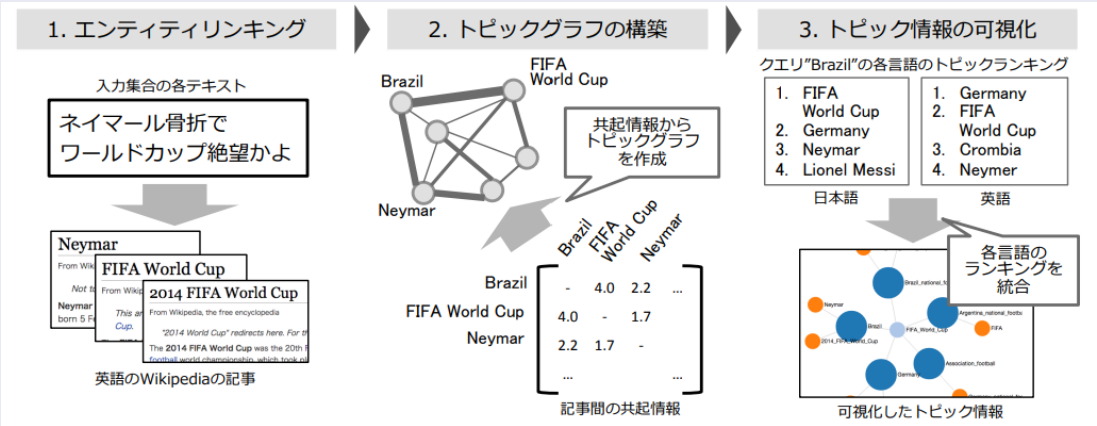

図 1: 試作システムの流れ

背景と目的

まとめ

試作システムの使用例

下の図では FIFA World Cup という英語の記事のクエリに対して英語, スペイン語, 日本語, アラビア語の 4ヶ国で共通して話題である記事が表示されている.

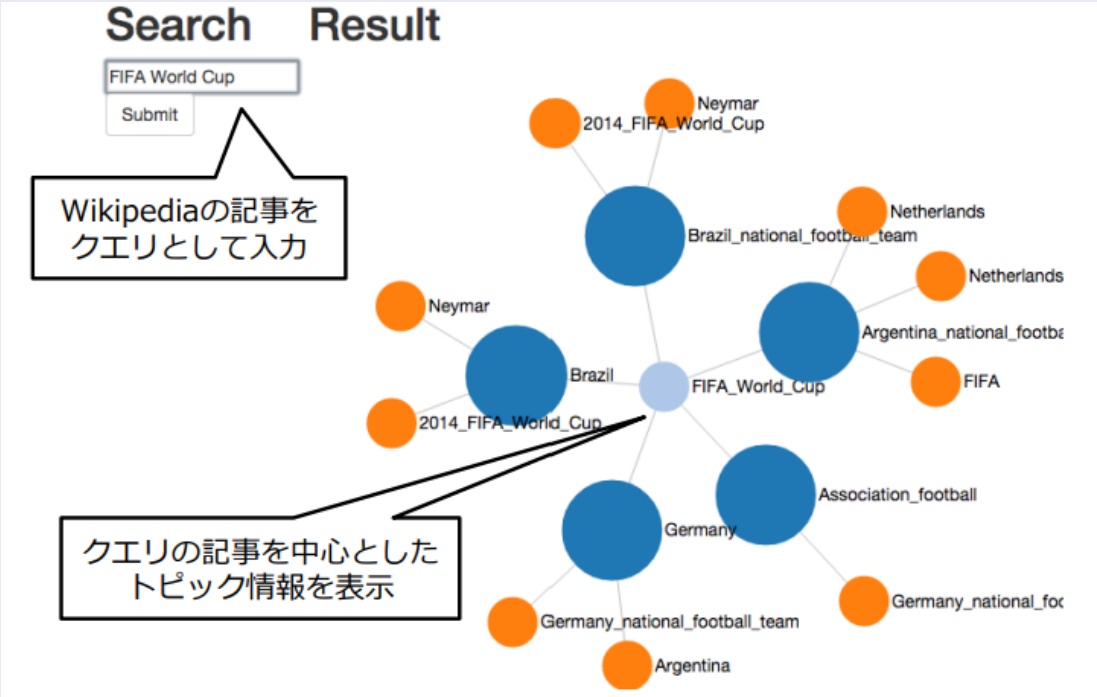

図 2: 使用例

手法について 1

背景と目的
まとめ

TAGME によるエンティティリンキング

エンティティリンキング: テキストの入力に対して, テキスト中に出現するエンティティを Wikipedia などのエントリに紐づけるタスクのこと.
TAGME は短文を対象とし, 識別子を使用しないエンティティリンキング手法
TAGME では, 入力テキストから Wikipedia のアンカーテキストとして用いられる語句をキーワードとして抽出し, キーワードから連想される記事の候補の中から互いに関連性の高い記事を付与するというシンプルな処理で高速高精度のエンティティリンキングを実現.

オリジナルの TAGME は単一言語のテキストを対象とした手法であるが, 本研究ではそれを Wikipedia の言語間リンクによって拡張した。
→オリジナルのものを適用し Wikipedia の記事を付与した後, その記事が英語の記事への言語観リンクを持っている場合は英語の記事に変換

手法について 2

背景と目的
まとめ

記事間関連度について

キーワードに関してはテキスト中に出現する Wikipedia のアンカーテキストすべてを抽出する. 記事の抽出では以下の式を用いる.

$$rel_a(p_a) = \sum_{b \in A \setminus \{a\}} \frac{\sum_{p_b \in p_g(b)} rel(p_b, p_a) \cdot Pr(p_b|b)}{|Pg(b)|b}$$

ここで $rel(p_b, p_a)$ は記事間関連度を, $Pr(p_b|b)$ はキーワード b がアンカーテキストとして使われる際に記事 p_b にリンクされる確率を表している.

手法について 3

背景と目的
まとめ

トピックグラフの構築

共起情報が記事間の関連性の強さを表しているとして, 記事をノード, 同一のテキストに付与された記事の共起回数をエッジとしたトピックグラフを構築した.

トピック情報の可視化

各言語におけるトピック情報の抽出には Affinity Propagation を適用した.
本研究では, エッジの重みとして共起回数を用いているため, ノード自身の重要度のスコアを実際にどれくらい言及されたかの回数, ノード間の関係を示すスコアを記事同士が関連して言及された度合として考える. これらのスコアを用いてランキングを作成し可視化する.

実験

背景と目的
まとめ

実験

4ヶ国語 (英語, スペイン語, 日本語, アラビア語) の Twitter ツイートを用いて予備実験を行った. 実験では 5 名の被験者に対して, 試作システムを用いて英語の Wikipedia の記事をクエリとした検索を自由に行ってもらいそれぞれの可視化手法についてのアンケートに答えてもらった.

アンケート内容

(1) すべての言語で共通して話題であるトピックの可視化について
(1-a) クエリに関連したトピックが表示されているか
(1-b) 可視化前の各言語のランキングと比較して, すべての言語で共通して話題であるトピックが表示されているか
(2) 特定の言語でのみ話題であるトピックの可視化について
(2-a) 上の (1-a) と同じ
(2-b) (2-a) の回答の理由がクエリに対して言語的な話題が抽出できていないかクエリに対して言語特有な話題がないのどちらか

背景と目的
まとめ

実験結果

すべての被験者が検索したクエリについて関連したトピックが表示されていたと回答した.
質問項目 2 に対してもほとんどの項目で平均値が 2 以上になった.
一方で言語に特有なトピックを抽出できていないという回答もあった.
このときに可視化されたトピックについて確認したところ言語に特有なトピックは表示されていたが, 話題になっていないトピックも表示されていた.

質問項目/被験者	A	B	C	D	E
(1-a)	3	3	3	3	3
(1-b)	2.4	2.8	2.7	3	3
(2-a) EN	1.6	3	2.6	1.6	2.5
(2-a) ES	2.6	3	3	3	3
(2-a) JA	2.3	3	3	3	2.5
(2-a) AR	2.2	3	3	3	2.5
(2-b) EN	0.7	-	-	2.5	-
(2-b) ES	-	-	-	-	-
(2-b) JA	1	-	-	-	-
(2-b) AR	2	-	-	-	-
(2-c) EN	2.5	2	2.4	2.7	2.5
(2-c) ES	2.8	1.8	2.6	2.8	2
(2-c) JA	2	2	2.2	2.2	2
(2-c) AR	3	2.2	2.8	2.8	2.5

図 3: アンケート結果

まとめ

背景と目的
まとめ

まとめ

多言語なソーシャルメディアにおけるトピック情報を言語横断的に抽出・可視化するシステムを試作した.
試作システムでは複数の言語で共通したトピックや特定の言語でのみ話題となっているトピックを可視化できていることを確認した.

課題

本実験の評価のデザインが 5 人に対するアンケートということであまり適切なものではなかったように思える.
エンティティリンキングの精度向上
詳細なトピック情報を同時に提示することで, より効率的な可視化.