DEIM2020 P1-15(day1 p80)

店舗の分散表現に対する意味演算を用いた飲食店検索手法

高橋　輝†　北山　大輔†

† 工学院大学情報学部システム数理学科　〒 163-8677 東京都新宿区西新宿 1 丁目 2 4 − 2
E-mail: †j316160@ns.kogakuin.ac.jp, ††kitayama@cc.kogakuin.ac.jp

あらまし　飲食店検索サイトを利用する際，ジャンル等の大まかな要求は容易に思いつくことができるが，具体的な料理名や味のバリエーションなどの細かい要求を適切に検索結果に反映することは難しい．そこで本研究では，既知の飲食店に対して任意のキーワードを足し引きすることにより，細かい要求を反映できる飲食店検索を提案する．そのために本稿では，飲食店とキーワードを横断的に扱える分散表現の検討を行った．具体的には学習済みの単語分散表現を用いた文書ベクトル作成手法である SWEM に基づき，店レビューサイトに投稿されたレビューを形態素解析することで名詞，動詞，形容詞の原型を抽出し各レビューの文書ベクトルを決定する．これら文書ベクトルの平均をとることで各飲食店のベクトルとする．このように得られる飲食店ベクトルとキーワードベクトルを用いた意味演算を検討し，検索システムを用いた従来手法との比較を行い評価実験を行った．

キーワード　飲食店検索, 分散表現, 単語の演算

1　は じ め に

近年，食べログやぐるなびなどの飲食店検索サイトが数多く存在している．このような飲食店検索サイトを利用する際，ジャンル等の大まかな要求は容易に思いつくことができるが，具体的な料理名や味のバリエーションなどの細かい要求を検索結果に反映させることは難しい．

例えば，ある飲食店 A に対し「もう少しさっぱりしたものが良い」などの細かい要求が発生するが,「飲食店 A さっぱり」のような検索では意図する結果は得られないと考えられる．そこで本研究では，ある店舗にキーワードを加減算することで，理想の要素を強調，または抑制した飲食店検索を可能にする手法を検討する．そのためには，飲食店とキーワードに四つ意味的演算をできるようなモデルを構築する必要がある．そこで，本稿では食の表現の多様さや表記揺れに強い方法として単語の分散表現を用いる．またそれを用いて飲食店の分散表現を得る．

単語と飲食店のベクトル空間を同一にできる可能性のある手法として，2.1 節で述べる SWEM がある．本項では SWEM によって適切に単語と飲食店を意味演算できることを示す．

本論文の構成を以下に示す．2 章で先行研究や関連研究の紹介をう．3 章にて提案方式の仔細を述べる．4 章にて検索システムの評価実験の概要をまとめ，結果から考察を述べる．5 章にてまとめを行う．

2　研究のアプローチ

2.1　SWEM

本研究では，店舗とキーワードを同じ特徴空間で表現できる必要がある．そこで文書の分散表現の生成手法として，Simple Word Embedding Based Models(SWEM) という手法を採用する [1]. この手法は単語の各次元の最大値や平均値を文書ベクトルとして採用する手法である．他の文章に対する固定次元の分散表現を得る手法としては，doc2vec [2] や Skip-thoughts [3] などが挙げられる．これらの手法はキーワードベクトルに加えて文章ベクトルを得るためのニューラルネットワーク自体を，大規模コーパスから学習させる必要があるが，SWEM は学習パラメータを必要とせず，計算コストも低い点で優れている．

2.2　研究の概要

本研究では店舗とキーワードとの演算を行うため，まずグルメレビューから単語を抽出し，SWEM を用いることでレビューの文書ベクトルを作成する．この時，絶対値が最大値となる値を採用する SWEM_absmax 法を定義する．また，これを用いて，店舗ごとにレビューのベクトルの平均を取ったものを飲食店ベクトルとする手法を提案する．それらをキーワードと演算することで細かい要求を満たした飲食店の検索を可能とする．

2.3　関 連 研 究

本研究で用いているレビューには，目的や理由などの特定の情報を抽出，有用性を判定する研究などが取り組まれている．倉橋ら [4] は Amazon レビューの「参考になったかどうか」の指標から有用性の判定実験を行った．有用でないレビューは店舗の分散表現には適していないと考えられ，これらのレビューを排除することで，本研究の精度も上がると考えられる．中山ら [5] はレビュー中に含まれる感情表現やその理由などが，テキストを理解，信用するための大きな要素であるということを明らかにした．このようにレビューには，その商品や店舗などの特徴を表す要素があると考え，本研究では分散表現を得るために用いている．

単語の分散表現手法として様々な技術が研究されている．水野ら [6] は楽曲の要素を足し引きする音楽推薦を行うために word2vec で学習させた単語ベクトルに IDF の重みを加えた手法と SCDV を用いた手法の 2 パターンを提案している．山

本ら [7] は Bag-of-Words と word2vec で提案されている skip-gram を組み合わせた単語ベクトルとあらかじめ単語をクラスタリングすることで，類似店舗を算出している．本田ら [8] は商品対の目的判定による商品推薦を行うために，商品カテゴリごとのレビューの単語頻度の χ 二乗値を算出し，各単語の特徴量を特徴ベクトルとして採用している．柳本ら [9] は文書の類似度を，分布の要素間の距離を考慮した距離を定義することができる Earth Mover's Distance を用いることで，単語間の類似性を考慮した文書類似度計算手法を提案している．本研究では食の表現の多様さに対応するため，字面の近しい単語同士により意味のまとまりをもたせることができる fasttext を用いている．

また文書の分散表現手法も様々な技術が研究されている．吉田ら [10] は主観的特徴の意味的演算による観光スポット検索システムを実装するため，ある観光スポットに対し投稿されたレビュー全てを一文とみて Paragraph Vector を用いて文書ベクトルを得ている．小中ら [11] は意味類似文の判定のため，ユークリッド距離やレーベンシュタイン距離を拡張することで，STS に基づく類似度手法を提案している．これらの手法では，計算コストが大規模な点から本研究では SWEM を採用している．

森田ら [12] は，嗜好とリアルタイム性を考慮した飲食店検索の研究を行なっている．この飲食店検索システムは，飲食店とシズルワードを蓄積することでユーザの細かい要求を反映できるような検索システムを設計している．これらの研究は，細かい要求を検索に反映させるという目的が本研究と類似しているが，本研究では，とある店舗から動的に検索結果が変化可能な点が異なっている．

3 提 案 手 法

本章では飲食店検索のための，レビューから飲食店の分散表現を得る手法とキーワードとの演算手法について述べる．レビューから飲食店の分散表現を得て，キーワードとの演算を行うまでの流れを図 1 に示す．

まず，グルメレビューの形態素解析を行う．次に，グルメレビューに対し fastText[1] を用いて単語の分散表現を得る．その後，個々のレビュー単位で SWEM 法によるレビューベクトルを得る．本研究で用いる SWEM_absmax については 3.2 節で述べる．各店舗におけるレビューベクトルの平均を飲食店ベクトルのレビュー部とし，カテゴリの考慮のためカテゴリ部を作成して結合し，飲食店ベクトルとする．飲食店ベクトル生成の流れを図 2 にまとめる．これをキーワードと演算することで演算後のベクトルを作成した．この演算後のベクトルと飲食店のベクトルの cos 類似度を取ることで，指定の要素が加減算された飲食店の類似度が高くなると考えられる．

3.1 キーワードベクトルの生成

まず MeCab [13] を用いて，飲食店検索サイトのグルメレビューの形態素解析を行う．MeCab とは，自然言語処理分野

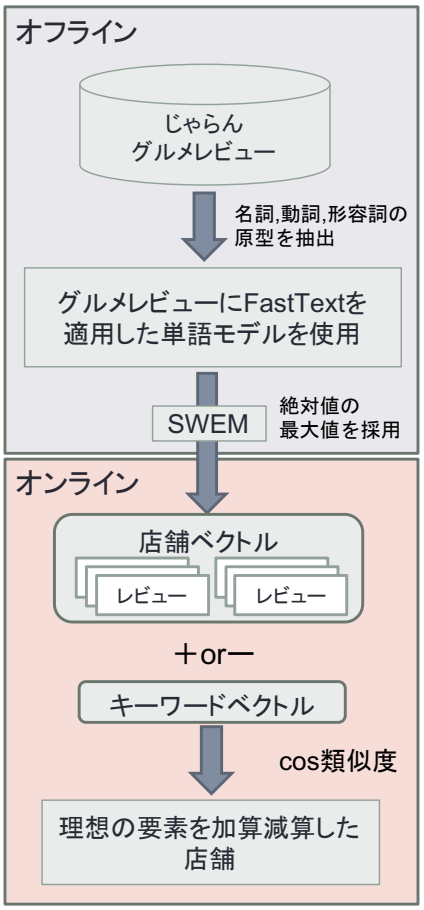

図 1 キーワード演算までの流れ

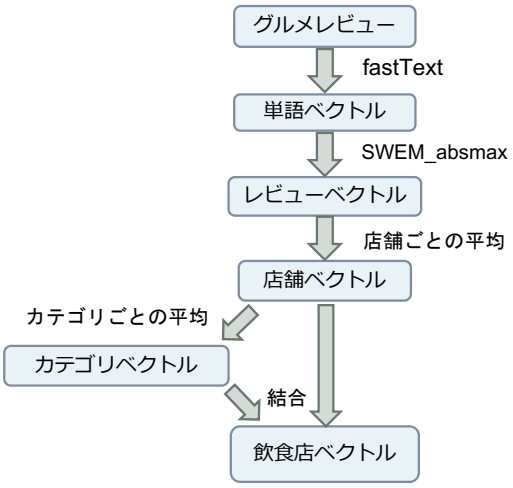

図 2 レビューから飲食店ベクトル生成までの流れ

で事前処理として用いられることが多い形態素解析エンジンである．その際，そのレビューの特徴を表すものを名詞，動詞，形容詞の 3 種類と仮定し，それらの原形を抽出する．辞書は NEologd[2] を使用した．NEologd を用いるのは週 2 回以上更新され，新語や固有表現に強い辞書であるため，レビューなど固有表現が出やすい文書にも対応できると考えたためである．抽出されたレビューの名詞，動詞，形容詞に対してじゃらん net[3]

1 : https://github.com/facebookresearch/fastText

2 : https://github.com/neologd/mecab-ipadic-neologd

3 : https://www.jalan.net/gourmet/

表 1 「あっさり」を入力した時のコーパスの比較

Wikipedia		グルメレビュー	
単語	類似度	単語	類似度
あっけなく	0.53	さっぱり	0.77
悔し	0.49	こってり	0.65
食い下がる	0.48	アッサリ	0.64
ピラフ一味	0.47	サッパリ	0.62
一蹴	0.46	コッテリ	0.60

表 2 fastText のパラメータ

パラメータ	設定値
dim	300
epoch	10
minCount	20

のグルメレビューをコーパスに fastText で学習した日本語モデルを用いて，300 次元のベクトルを得る．通常 fastText のコーパスは Wikipedia が一般的に採用されるが，Wikipedia では一般的な単語の分散表現となってしまいグルメに適した意味的演算には不向きである．実際に Wikipedia で学習させた時とグルメレビューで学習させた時の「あっさり」という単語の類似単語を類似度が高い順に上位 5 件抽出し，比較した．その結果を表 1 に示す．表 1 を見てみると，Wikipedia で学習させた場合，いずれの単語もグルメとは関係しないような単語が上位に来てしまっている．対してグルメレビューで学習させた場合は，グルメドメインとして意味が近しいものが上位に現れている．また，今回 fastText で学習させる上で使用したパラメータを表 2 に示す．

3.2 SWEM_absmax を用いたレビューベクトルの生成

飲食店のベクトルは，その飲食店の全レビューベクトルの平均で生成する．本節ではレビューごとの文書ベクトルの作成を説明する．レビューベクトルは，3.1 節のキーワードベクトルを用いた SWEM で作成する．先行研究である SWEM はレビューの各キーワードベクトルの最大値を採用した SWEM_max や，平均値を採用した SWEM_mean を文書ベクトルとする手法である．SWEM_max は負の方向に大きいという特徴は扱われておらず，負の方向にも特徴を持つキーワードベクトルとの演算に対して，不適切であると考える．一方 SWEM_mean は，ノルムが小さくなる傾向があり，また，文中の単語で特徴を打ち消してしまうことが考えられる．そこで本研究では，正に大きい値だけでなく，負に大きい値でもその単語の特徴となると考え，各次元で絶対値の最大値をとる値を用いる SWEM_absmax 法を定義する．

v をあるレビューに含まれる単語ベクトルとし L をそのレビューの単語数とすると，各ベクトルの最小値で生成するベクトル，最大値で生成されるベクトルはそれぞれ式 1，式 2 で定義できる

$$r_i{}^{min} = Min - pooling(v_1, v_2, v_3, ..., v_L) \quad (1)$$

$$r_i{}^{max} = Max - pooling(v_1, v_2, v_3, ..., v_L) \quad (2)$$

これらを用いて，ある店舗の i 番目のレビューの SWEM_absmax を式 4 で定義する

$$r_i = [absmax(r_i{}^{min}{}_1, r_i{}^{max}{}_1), ..., \quad (3)$$

$$absmax(r_i{}^{min}{}_D, r_i{}^{max}{}_D)] \quad (4)$$

ここで $r_i{}^{min}{}_1$，$r_i{}^{max}{}_1$ はそれぞれのベクトルの 1 番目の要素であり，D は次元数である．$absmax$ は引数のうち絶対値が最大である値を返す関数である．

3.3 飲食店ベクトル

飲食店ベクトルは，有するレビューの平均ベクトル 300 次元 (レビュー部) と属するカテゴリベクトル 300 次元 (カテゴリ部) の計 600 次元で表現する．これを以降 Mean_SWEM_absmax と定義する．各飲食店に付随している 62 種類のカテゴリをもとに，各カテゴリに属する飲食店のレビュー部の平均を取ったものをカテゴリ部のベクトルとする．それぞれ，レビュー部はレビュー部のノルム，カテゴリ部はカテゴリ部のノルムで割ることで正規化する．

3.4 キーワードとの演算

SWEM で作られた飲食店ベクトルと元となるキーワードベクトルは原理的に同じ空間であり，そのため直接の演算が可能であると考えられる．飲食店とキーワードとの演算は，飲食店のベクトルからキーワードベクトルを加減算することで実装する．こうすることで「ミスタードーナツ + 高い = クリスピークリームドーナツ」といった意味的演算を可能とし，より直感的な検索が可能になることを期待している．

具体的な演算方法としては，飲食店ベクトルと自身のノルムで正規化したキーワードベクトルで，600 次元同士の加減算を行う．キーワードベクトルは 300 次元のため，カテゴリ部を零ベクトルとして計算する．こうしてできたベクトルに対し，飲食店ベクトルとの cos 類似度を求めることで，指定のキーワードが足し引きされた店舗を検索する．

3.5 プロトタイプシステムの設計

検索システムとして，図 3 のようなものを作成する．行きたい飲食店が思い浮かばない状況を想定しているので，まずはランダムに出力された店舗群からユーザに気になる店舗を 1 件選択してもらう．店舗を選択すると加算減算が可能な入力部が出現する．この時，何か追加したい要素があれば入力部に単語を入力する．逆に削除したい要素があるときはキーワード群から選択する．キーワード群は類似単語をみることができ，類似単語内には対義語と見られるものも存在する．もし対義語と見られるものが類似単語内に存在した場合，対義語と見られるものを加算部に入力した状態で，引きたいキーワードを減算部に入力してもらう．こうしてできた演算後のベクトルにより，検索結果として類似度上位の店舗が出力される．

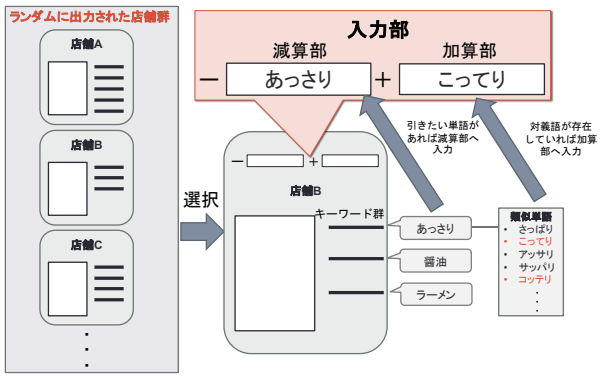

図 3　検索システムの構造

4　実　　験

4.1　実 験 方 法

ベクトル生成の妥当性を評価するため提案手法と比較手法に対し MAP で精度を評価する．比較手法にはカテゴリベクトルの有無，レビューベクトルの平均を店舗ベクトルとして採用するかどうか，SWEM を絶対値の最大値（absmax），最大値（max），平均値（mean のそれぞれを採用する計 11 手法を用意した．これらの手法を用いて，著者らが選定した飲食店に対し加算と減算それぞれ 3 クエリずつ行い，それぞれの結果に対し MAP で評価を行う．具体的には以下の 6 種のクエリを用いた．

- かつや ＋ ソース
- ラーメン次郎 ＋ ヘルシー
- 吉野家 ＋ 女性
- ゴディバ - 高い
- 蒙古タンメン - 辛い
- ぎょうざ 松龍軒 - にんにく

またデータセットには飲食店検索サイト「じゃらんグルメ」から 474,457 件のレビュー，店舗にして 76,703 件を用いてベクトルを作成した．

正解データの作成には，提案手法比較手法を用いた加算と減算それぞれ 3 クエリに対し，それぞれ類似度が高い上位 10 件の飲食店の集合を用意する．この際，入力店舗が検索結果に含まれた場合は削除している．また，データセットの都合上，現在存在しない店舗についても削除した．そのため，各検索結果は，10 件に満たないものもある．次に，被験者に検索意図を提示し，それまでの飲食店が適切な結果であるかどうかを判定してもらう．判断する上で，じゃらんの各飲食店の URL を付随した．被験者はクラウドソーシングサービスである CrowdWorks を用いて，1 クエリにつき 5 人ずつ集め判定してもらい，多数決で正解を決定した．

4.2　実験結果と考察

各 12 手法を用いて MAP を算出した結果を表 3 に示す．表 3 の列はカテゴリベクトルの有無を表しており，上からから SWEM の最大値を採用する手法，平均値を採用する手法，絶対値の最大値を採用する手法の順に並んでおり，4 行目からは店舗ベクトルをレビューベクトルの平均を採用する手法に加え，SWEM の各手法が先ほどと同じ順に並んでいる．

表 3　各手法での MAP

	カテゴリなし	カテゴリあり
SWEM_max	0.2680	0.3061
SWEM_mean	0.3920	0.7191
SWEM_absmax	0.2917	0.5948
Mean_SWEM_max	0.2515	0.2553
Mean_SWEM_mean	0.4128	0.7222
Mean_SWEM_absmax	0.6396	0.9140

表 4　カテゴリありでの検索結果例

正解	店舗	カテゴリ
○	横浜中華街 皇朝レストラン	28
○	横浜大飯店	28
○	中国料理 金紗沙	28
○	宇都宮みんみん 駅東口店	28
○	餃子の店　山女	28
○	第 7 ギョーザの店	28
○	餃子の王将 那珂川店	28

まずカテゴリ有無の差異について見てみると，カテゴリベクトルありの方が概ね MAP が高い結果となっている．実際に Mean_SWEM_absmax 法での「ぎょうざ 松龍軒 - にんにく」の検索結果上位のカテゴリありの場合を表 4 に，カテゴリなしの場合を表 5 に示す．表 4 をみると，上位 7 件全てが元店舗と同じカテゴリ ID になっていることがわかる．しかし表 5 ではほとんどの店舗のカテゴリがバラバラになっていることがわかる．この結果からカテゴリベクトルの有効性が確認できる．

次にレビューの平均を店舗ベクトルとして採用した際の差異を見てみると，こちらも概ねレビューの平均を用いた手法が，全レビューを一文としそれを店舗のベクトルとしてそのまま採用する手法よりも上回る結果となった．これはレビューを平均化せずに店舗ベクトルとして採用したため，ノイズのような単語が強く効きすぎてしまい適切な検索が得られなかったと考えられる．

最後に，SWEM の最大値，平均値，絶対値の最大値を採用する手法の差異を比較していく．提案手法に比べ最大値を採用した SWEM_max，平均値を採用した SWEM_mean 共に MAP は低い結果となっている．SWEM_max は負の値を考慮しないためベクトル表現の幅が狭くなり，適切に検索意図を反映できなかったことが原因だと考えられる．また SWEM_mean は平均化したためベクトルに特徴が反映されづらく，SWEM_absmax の方がより特徴を反映した検索が可能だったと考えられる．

5　ま　と　め

飲食店検索サイトを利用する際，細かい要求を適切に検索結果に反映することは難しい．そのため本研究では，既知の飲食店に対して任意のキーワードを足し引きすることにより，細かい要求を反映できる飲食店検索を提案した．提案手法の

表 5　カテゴリなしでの検索結果例

正解	店舗	カテゴリ
×	伊予のご馳走　おいでん家	4
×	郷土料理 五志喜	6
○	横浜中華街 皇朝レストラン	28
×	伊豫水軍	4
×	ほづみ亭	4
×	道後麦酒館	2

Mean_SWEM_max 法では，キーワードベクトルの絶対値の最大値をとる値をレビューベクトルとして採用し，それらの各レビューの平均を飲食店の分散表現として得た．これによりキーワードと飲食店との同一空間での検索を可能にし, キーワードにより詳細な特徴を指定することができる．実験では提案手法を用いることで，比較手法より適切に検索ができる可能性を確認した．

今後の課題としては，店舗の特徴キーワードの自動抽出を行い，選択店舗に対するクエリ生成支援を含むプロトタイプシステムを構築し，評価することがあげられる．

謝　　辞

本研究の一部は，2019 年度科研費基盤研究 (C)(課題番号：18K11551) によるものです．ここに記して謝意を表すものとします

文　　献

[1] Dinghan Shen, Guoyin Wang, Wenlin Wang, Martin Renqiang Min, Qinliang Su, Yizhe Zhang, Chunyuan Li, Ricardo Henao, and Lawrence Carin. Baseline needs more love: On simple word-embedding-based models and associated pooling mechanisms. *CoRR*, Vol. abs/1805.09843, pp. 1–13, 2018.

[2] Quoc V. Le and Tomas Mikolov. Distributed representations of sentences and documents. *CoRR*, Vol. abs/1405.4053, pp. 1–9, 2014.

[3] Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Skip-thought vectors. *CoRR*, Vol. abs/1506.06726, pp. 1–9, 2015.

[4] 倉橋宏幸, 青野雅樹. Amazon レビューを用いた有用性の判定実験. 情報科学技術フォーラム講演論文集, 第 12 巻, pp. 101–102, aug 2013.

[5] 中山記男, 神門典子. レビューにおける「理由」の分析　～被験者実験より～. 情報処理学会研究報告自然言語処理（NL）, 第 1 巻, pp. 81–88, jan 2006.

[6] 水野智公, 亀谷由隆. 単語のベクトル表現に基づき楽曲要素の足し引きを行う音楽推薦. 情報処理学会第 81 回全国大会講演論文集, No. 1, pp. 345–346, feb 2019.

[7] 山本真史, 山崎俊彦, 相澤清晴. Bag of words と skip-gram 併用によるレビュー・店舗間類似度評価とそれに基づく店舗推薦. 情報処理学会第 78 回全国大会講演論文集, No. 1, pp. 543–544, mar 2016.

[8] 本田達也, 北山大輔, 角谷和俊. オンラインショッピングサイトにおけるレビューを用いた商品対の目的判定による商品推薦. 平成 24 年度 情報処理学会関西支部 支部大会 講演論文集, pp. 1–5, 2012.

[9] 柳本豪一. 単語の分散表現を利用した文書類似度. 人工知能学会全国大会論文集, Vol. JSAI2015, pp. 1–2, 2015.

[10] 吉田朋史, 北山大輔, 中島伸介, 角谷和俊. ユーザレビューの分散表現を用いた主観的特徴の意味演算による観光スポット検索システム. 第 9 回データ工学と情報マネジメントに関するフォーラム, pp. 1–5, mar 2017.

[11] 小中史人, 三浦孝夫. 語の並びを考慮した意味類似度手法の提案. 第 7 回データ工学と情報マネジメントに関するフォーラム, sep 2015.

[12] 森田真季, 宮部真衣, 荒牧英治, 灘本明代, 吉野孝. 嗜好とリアルタイム性を考慮した飲食店検索システムの構築. 2018 年度 情報処理学会関西支部 支部大会 講演論文集, pp. 1–3, sep 2018.

[13] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying conditional random fields to Japanese morphological analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pp. 220–237, Barcelona, Spain, July 2004. Association for Computational Linguistics.