検索語の関連語彙と頻出フレーズを用いた検索意図明確化手法

廖　　子揚† 　田島　敬史†

† 京都大学情報学研究科　〒 606–8501 京都市左京区吉田本町 36-1
E-mail: †liao@dl.soc.i.kyoto-u.ac.jp, ††tajima@i.kyoto-u.ac.jp

あらまし　多くの検索エンジンでは，ユーザの検索意図は複数のキーワードとフレーズで構成されるクエリで表されるが，そのようなクエリでは複数の解釈が可能なあいまいなクエリが多数存在する．多くの検索エンジンは，クエリログデータを用いて生成したクエリの候補をユーザに提示することで，適切なクエリを見つけるのを手助けするが，この手法ではマイナーなクエリ意図に対しては適切なクエリを推薦できない．本研究では，与えられた複数検索語の間の関係を表しうるフレーズを Web コーパスから抽出する．また，その際，マイナーな検索意図にも対応するため，検索語をより一般的な語に置き換えたものも用いる．そして，抽出したフレーズを元のクエリに追加したものをクエリ候補としてユーザに提示し，ユーザは自らのクエリ意図に合うフレーズを選択する．これにより，検索語間の関係がより明確なクエリとなることが期待できる．

キーワード　情報推薦，クエリ拡張，意図の曖昧さ解消

1. 概　　要

World Wide Web の急速な発展に伴い，検索エンジンはユーザにインターネット上で必要な情報を検索するためのかけがえのない方法を提供する．現在の商用検索エンジンは，最も関連性の高い Web ページをユーザに推薦するのに成功しているが，検索結果の品質を低下させる可能性があるいくつかの顕著な問題がある．

まずは，自然言語に共通して存在する曖昧さの問題である．曖昧な用語を含むクエリは，検索エンジンをユーザの情報要求を満たさない Web ページを取得することでユーザを混乱させる可能性がある．また，[8] に報告されているように，ユーザはほとんど 1 つまたは 2 つの用語からなる短いクエリを使用する傾向があり，短いクエリは長いクエリより曖昧である可能性が高い．2006 年に 3ヶ月以上にわたって記録された商用検索エンジンのクエリログを分析すると，Web クエリの 19．4 %が 1 単語クエリであり，さらに Web クエリの 30．5 %が 2 単語クエリであることがわかった．そして，ほとんどの場合，ユーザが検索する理由は，検索しているトピックに関する知識がほとんどまたはまったくないということである．満足のいく回答を見つけるために，ユーザは何度も自分のクエリを変更する必要がある．

これらの問題を解決するために，Google(注1)，Bing(注2)，Yahoo(注3)などの多くの有名な検索エンジンは，クエリの推薦を行ってきた．これらは，推薦するクエリを生成するための情報としてクエリログを使用する．しかし，マイナーなクエリ意図に対する推薦はクエリログの中に情報が少ないため困難である．

本論文では，クエリ意図の曖昧さ解消手法を紹介する．本論文では 2 語のクエリのみを対象とする．例えば,「りんご 肥料」のようなクエリは,「りんごの肥料を探す」や「りんごを肥料にする方法」など複数解釈がある．クエリログの情報だと「りんごの肥料を探す」という意図が一般的なので，それに関するクエリは推薦の中で高くランクされる．しかし,「リンゴを肥料にする方法」に関する情報を探したいユーザにとっては，既存の検索エンジンのクエリ推薦から適切なクエリを見つけることが困難である．この問題を解決するために，我々は，Web の情報に基づいてクエリの提案を行い，マイナー意図のクエリでもをユーザに推薦することができる手法を提案する．

提案手法では，まず，クエリのすべての単語について，語彙意味関係を持つ下位語，同義語，上位語の単語をクエリをを収集し，これらで元の語を置き換えたクエリを収集する．これは，本研究の目的がマイナーな意図のための推薦であり，そのためには同義語や上位語も使って再現率を上げることが重要であるためである．元の語を置き換えたクエリの間の関係を表すフレーズを Web から抽出し，それを元のクエリに追加することで，新たなクエリを生成し推薦する．たとえばリンゴの場合，果物やバナナなどに拡張し，拡張後の単語集合を使用し，検索語間の関係をより多く見つけることができる．Web コーパスに「りんごは肥料として使う」という情報があれば,「りんご」と「肥料」の関係を抽出し，それをユーザに推薦することができる．検索エンジンは,「りんごは肥料として使用できる」と関連する情報を推薦できるので，ユーザは必要な情報を簡単に見つけることができる．次に類似する推薦クエリをクラスタリングし，同じ意味の結果を減らし，より多様性のある結果を生成する．クラスタリングは意味類似度に基づき，Affinity Propagation クラスタリングアルゴリズムを用いて意味の近い結果のクラスタリングを行う．

本稿の構成は以下のとおりである．2 節では，関連研究を紹介する．3 節では，提案手法について詳細に述べる．4 節では，提案手法に関する実験と評価について述べる．5 節では，まと

(注1)：http://www.google.com
(注2)：http://www.bing.com
(注3)：http://www.yahoo.com

めと今後の課題について述べる．

2. 関 連 研 究

本節では，本研究と関連する研究について言及し，本研究の位置付けについて述べる．クエリの曖昧さを解消する方法として，クエリ推薦が挙げられる．多数の検索エンジンはクエリ推薦機能を提供していて，この分野において多くの関連研究がある．ユーザはクエリ推薦機能を利用し，より自分の目的に合うクエリを見つけることができる．

2.1 Query Suggestion using Query Logs

現代のウェブ検索エンジンの大部分は，ユーザがクリックした URL や他のセッションデータなどのユーザの検索行動を記録する．Baeza-Yates ら [1] は，クエリログから抽出されたデータに対するクラスタリングプロセスに基づいて関連クエリを提案する方法を提案した．クラスタリングプロセスは，検索エンジンのクエリログに登録されたユーザの履歴プリファレンス上のコンテンツを利用する．このアプローチは，関連するクエリを検出するだけでなく，関連度に基づいてランキングも行う．

クエリログには，クリックスルーおよびセッションデータも含まれる．Cao ら [4] は，2 つのステップである新しいコンテキスト認識クエリ提案手法を提案した．オフラインモデル学習ステップでは，データの希薄さを解消するために，クリックスルーバイパータイトをクラスタリングすることによって，クエリが概念に要約される．セッションデータから，概念シーケンスサフィックスツリーがクエリ提案モデルとして構築される．ユーザの検索コンテキストは，ユーザによって提示されたクエリシーケンスを一連の概念にマッピングすることによって取得される．概念シーケンスサフィックスツリー内のコンテキストをルックアップすることにより，それらのアプローチは，コンテキスト認識方式でユーザに問い合わせることを使用している．

単一の検索セッションでは，ユーザはしばしば関連情報を得るために初期クエリを数回修正する．Jones ら [9] は，クエリの修正履歴を用いて，クエリ推薦を行う．Li ら [10] は，検索セッション中に共起するクエリが関連しているという仮説に基づき，クエリ間のペアワイズ類似度スコアを計算する方法を記述している．ユーザクエリが与えられると，このようにして識別された最も類似したクエリが提案として使用される．Cucerzan と White [5] は，クエリ推薦を生成するためにユーザランディングページ（ユーザがポストクエリナビゲーションの後に最終的に辿り着いたページ）を利用する．ユーザが送信したクエリの各ランディングページについて，これらのランディングページを上位 10 の結果の 1 つとして持つクエリログからクエリを抽出する．これらのクエリは，クエリ推薦として使用する．

Ma ら [11] は，特定のユーザが作成したクエリに対し，2 つのグラフ（ユーザ-クエリとクエリ-URL）の形でクリックスルーデータをマイニングし，効率的な two-level クエリ提案モデルを提案した．これをもとに，2 つのグラフを利用して低ランククエリの潜在的な特徴空間を学習し，次にそのアウトラインの特徴に基づくクエリ類似度グラフを構築する共同行列因子分解法を提案した．その後，彼らはオンラインランキングアルゴリズムを開発して，クエリ類似性グラフに見られる類似性を計算する．最後に，ユーザに意味的に関連する潜在的なクエリを推薦する．

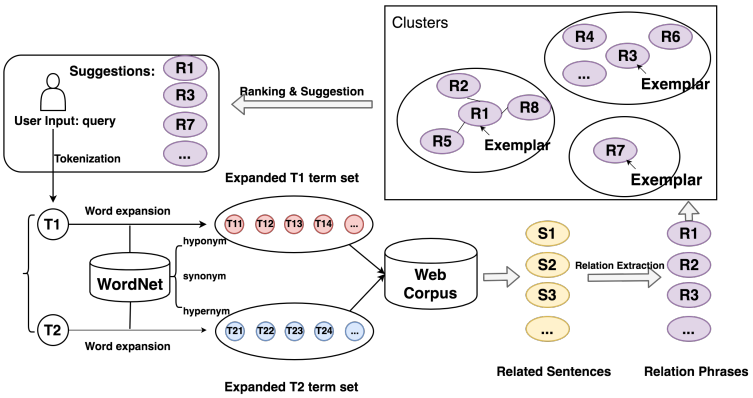

図 1 手 法 概 要

2.2 Query Suggestion without Query Logs

クエリログが主要なリソースとして使用していない研究の中では，Feuer ら [6] は，初期クエリが非常に限定的（非常に少数の結果）または広すぎる（多数の検索結果）場合に代替クエリを提案する Proximity search based system を提案した．曖昧なクエリの場合では，コーパス内のクエリ用語のすぐ近くに現れる最も頻繁な用語がクエリに追加され，検索結果が絞り込まれる．曖昧でないクエリの場合では，クエリの最も頻繁な部分フレーズが可能な推薦として表示される．

Bast と Weber [2] は CompleteSearch を開発した．これはユーザに複数の複雑な機能を提供する対話型検索エンジンである．ユーザがクエリを入力し，ユーザがクエリを変更したときにリアルタイムで結果をリフレッシュすると，自動補完機能が提供される．クエリ提案の結果は，検索結果から抽出され，その頻度の順に結果がランキングされる．CompleteSearch は，不完全なクエリに関連する完全なクエリに補完するのではなく，最後のクエリ単語を完成させるという形になるので限界がある．

Bhatia ら [3] は，教師なしでコーパスからのいくつかの関連するフレーズを組み合わせることによって形成された提案のクエリに限定されているアプローチを提供している．当然ながら，彼らのシステムは，結果が得られるかもしれないが，検索クエリとして使用されているにもかかわらず，検索クエリとして意味がない推薦が生成される．

3. アプローチ

このセクションでは，Web コーパスにある情報をどのように使用してクエリの意図を明確にするかを紹介する．まず，より一般的な結果を取得するためにクエリ拡張し，Web コーパスの情報から拡張後のクエリに関連する文書を取得する．次に，関連する文書からの単語間の関係フレーズを抽出し，意味的類似度に基づいてこれらの結果をクラスタリングする．最後に，ランキングの上位 10 件の結果をユーザに推薦する．

3.1 クエリ拡張

例えば，「りんご 肥料」の例では，りんごと肥料だけを使っ

表 1　実験に使用したクエリ

id	query	common search intent	minor search intent
1-1	apple fertilizer	Information about fertilizer for apple	How to change apple into fertilizer
1-2	banana fertilizer	Information about fertilizer for banana	How to change banana into fertilizer
1-3	orange fertilizer	Information about fertilizer for orange	How to change orange into fertilizer
2-1	Kyoto bank	Information about the Kyoto Bank	Information about the bank in Kyoto
2-2	Japan bank	Information about the Bank of Japan	Information about the bank in Japan
2-3	China bank	Information about the China Bank	Information about the bank in China

表 2　Phrase cluster of “orange fertilizer”

cluster center	cluster phrases
is	is
	found in
	has
provides	contains
	offers
	provides
is runoff from	is runoff from
	are contaminated with
	finds a new home on
contain amounts of	contain amounts of
was a 50-50 mix of	was a 50-50 mix of

てウェブから関連情報を探すと，この組み合わせに正確に一致する情報が限られているため，結果はほとんど得られない．クエリ用語が一般的に使用されていない場合，ウェブから情報を取得することは難しく，したがって，いくつかであるの関連フレーズを取得することはできない．関係フレーズを取得する際に Web コーパスを使用する前に，より一般的な用語を得るためにクエリ用語を拡張する．ここでは wordsapi(注4) と呼ばれる Web API を使用し，この api は主に WordNet [12] にあるデータを使用し，同義語，上位語と下位語の関連語を取得する．ユーザが入力した 2 語クエリ $query = \{a_0, b_0\}$ で表すことができ，a と b はクエリに含まれている単語である．API を用いて，a と b の同義語，上位語と下位語関係を持つ単語を取得する．このようにして生成された単語集合を $A = \{a_0, a_1, a_2, ..., a_n\}$, $B = \{b_0, b_1, b_2, ..., b_n\}$ とする．

3.2　関 係 抽 出

拡張後のクエリを使用し，web コーパスから 2 つの単語に関連フレーズを抽出する．ある文から派生する語句を抽出するために，この文書の中の二項関係を抽出することができる Open Information Extraction（Open IE）[14] を使用する．Open IE が抽出した関係は（A,B,C）のようなトリプルで表現できる．A と C は単語で，B は A と C の間の関係を表す．たとえば「The U.S. president Barack Obama gave his speech on Tuesday to thousands of people.」のような文章があり， Open IE は，(Barack Obama, is the president of, the U.S.) のようなバイナリ関係を得ることができる．この例では，「Barack Obama」と「the U.S.」の関係フレーズを抽出することができる．拡張後のクエリについては，Open IE に基づいて Web コーパス内の文章を解析し，クエリ単語間に可能な関係フレーズを得ることができる．まず， 拡張後のクエリ単語集合を用いて， 新たなクエリを生成する．$A = \{a_0, a_1, a_2, ..., a_i\}$ と $B = \{b_0, b_1, b_2, ..., b_j\}$ の単語をそれぞれ組み合わせをし， クエリ集合 $Q = \{\{a_0, b_0\}, \{a_0, b_1\}, \{a_0, b_2\}, ..., \{a_i, b_j\}\}$ を生成する．それから， Open IE と ClueWeb のデータを利用し， クエリ集合 Q 中のクエリに関する文書を抽出し， クエリ単語の間の関係フレーズを抽出する．抽出したフレーズの集合を $F = \{\{f_1, feq_1\}, \{f_2, feq_2\}, ..., \{f_n, feq_n\}\}$ とする．f は抽出したフレーズであり， feq は f が抽出したフレーズ集合の中の出現頻度を表す．

3.3　クラスタリング

Web コーパスから多くの関連フレーズを得ることができるが，フレーズが多すぎる可能性があり，いくつかのフレーズが同じ意味を持つ可能性があるため，すべての結果をユーザに推薦することは困難である．より多様性がある結果をユーザに推薦するために，意味類似度に基づいて関係フレーズをクラスタリングする．類似度を計算するために，Mueller ら [13] が提示した文の類似度を計算する手法を使用する．これは，可変長シーケンスの対からなるラベル付きデータのための Long Short-Term Memory（LSTM）ネットワークのサイアム適応を提示し，符号化するために固定サイズのベクトルを使用する LSTM に同義語情報を補足したワード埋め込みベクトルを提供する．

意味的類似度を計算した後，類似度の高い関連フレーズをクラスタリングし，結果の数を最小限に抑える．ここでは，Affinity Propagation (注5) クラスタリングアルゴリズムを使用する．Affinity Propagation はデータポイントの間でメッセージと呼ばれる値を再帰的に収束するまで計算する．そして収束したメッセージの値から exemplar と呼ばれるクラスタを代表するデータを求める．Affinity Propagation は今までの k-means 法 [7] などのクラスタリング手法と比較し，（1）クラスタリングの誤差が少ない，（2）非対称で三角不等式の成り立たない類似度に対応する，（3）クラスタリング結果が初期状態に依存しないなどのメリットがあるという利点がある．Affinity Propagation を用いて F をクラスタリングしてできたクラスタの集合を $C = \{c_1, c_2, ..., c_n\}$ で表す．c_i には意味的類似度の高いフレーズが入っている．c_i の exemplar となるフレーズは e_i で表す．

3.4　ランキングアルゴリズム

関係フレーズをクラスタリングした後，結果をランキングし，

（注4）：https://www.wordsapi.com/

（注5）：http://scikit-learn.org/stable/modules/clustering.html

表 3　Query suggestion results of Bing and our approach

Query = apple fertilizer		Query = orange fertilizer	
Bing	Our approach	Bing	Our approach
best apple tree fertilizer	apple is fertilizer	atlas fertilizer	orange is fertilizer
sugar apple fertilizer	apple wont grow fertilizer	hengam fertilizer	orange provides fertilizer
atlas fertilizer	apple has about fertilizer	mosaic fertilizer	orange is runoff from fertilizer
hengam fertilizer		cru fertilizer	orange contains amounts of fertilizer
mosaic fertilizer		international fertilizer association	orange was a 50-50 mix of fertilizer
cru fertilizer		atlas fertilizer corporation	
international fertilizer association		fertilizer company	
atlas fertilizer corporation		fertilizer production	
Query = banana fertilizer		Query = Kyoto bank	
Bing	Our approach	Bing	Our approach
banana fertilizer 15 5 30	banana had been grown without fertilizer	the bank of kyoto ltd	Kyoto is located in bank
banana fertilizer requirement	banana purchase fertilizer	kyoto currency exchange	Kyoto be on bank
organic banana fertilizer	banana contains fertilizer	kyoto images	Kyoto was bank
fertilizer for banana trees		kyoto convention	Kyoto is home to bank
how to make banana fertilizer		kyoto photography	Kyoto had bank
banana tree fertilizer 6 2 12 formula		citibank atm in japan	Kyoto lies at bank
banana fertilizer for the garden		fertilizer company	Kyoto lit bank
banana fertilizer 6 2 12		fertilizer production	Kyoto sits on bank
			Kyoto built on bank
			Kyoto set beautifully on bank
Query = Japan bank		Query = China bank	
Bing	Our approach	Bing	Our approach
largest banks in japan	Japan is on bank	china bank philippines	China made bank
japan bank holidays	Japan are bank	china banking online	China is on bank
banks in tokyo japan	Japan promulgated bank	bank of china personal banking	China has raised bank
bank of japan policy rate	Japan are founded by bank	china bank branches	China reduced bank
big banks in japan	Japan have bank	people's bank of china website	China are bonded to bank
list of japanese banks	Japan is buying bank	china bank philippines website	China is located on bank
japanese banking	Japan is declared bank	bank of china internet banking	China are protected on bank
boj rates	Japan are very important to bank	icbc bank china online banking	China is the part of bank
	Japan is located on bank		China has been declared bank
	Japan generously bank		China has been bank

クエリ推薦を行う．この場合，クラスタ内のフレーズ頻度の合計を標本のスコアとして使用する．上位 10 件の結果を得て，拡張されたクエリ用語の代わりに元のクエリ用語に接続し，提案クエリを作成する．c_i の exemplar となるフレーズ e_i のスコアは，c_i の中のフレーズの出現頻度の和である．式 (1) はスコアの計算式である．

$$Score_{e_i} = \sum_{f_j in c_i} feq_j \quad (1)$$

4. 評価実験

本節では，提案手法を用いて行った実験とその評価について述べ，結果についての考察を行う．本実験の目的は提案手法の有効性について検証することである．複数の解釈可能なクエリを用意し，クエリ推薦結果を Bing 検索エンジンと比較した．マイナーな検索意図に対する提案されたクエリの検索結果の精度によって評価する．表 1 は，実験用のクエリと解釈可能な意図を示す．

表 4　Precision of minor search intent

id	1-1	1-2	1-3	2-1	2-2	2-3
Our approach	0.03	0.23	0.02	0.14	0.19	0.07
Bing	0	0.24	0	0.03	0.31	0

4.1 Test Queries

表 1 のクエリを用いて，本研究の手法を評価する．ここでは，複数解釈可能なクエリを選択する．例えば，“apple fertilizer” をクエリとして検索エンジニに入力すると，メインの検索意図は，“リンゴに使用する肥料に関する情報を探したい” である．一章で論述したように，“apple fertilizer” の検索結果がほとんどリンゴの肥料に関するページであり，“リンゴを肥料にする方法” を探したいユーザは自分のクエリを変更する必要がある．“apple fertilizer” のようなクエリを 6 個を用意し，提案手法を評価する．

4.2 Query Expansion

WordApi(注6) を使用し，クエリの単語を拡張する．“orange” については，語彙意味論関係を使ってそれを拡張することができる．

4.3 Relation Phrase Clustering

クエリ単語を拡張した後，これらの単語を使用し，2 つのクエリ単語間の関係を表すフレーズを Web コーパスから抽出する．表 2 にクラスタ結果を示す．

4.4 Query Suggestion Generation

関係フレーズクラスタリングの後，クラスタセンターを使用してクラスタを表し，それを元のクエリに接続して新しいクエリを生成する．表 3 は，Bing のクエリ推薦の結果と，提案手法によって作成されたクエリを示している．

(注6)：www.wordsapi.com

4.5 Result

表 4 に実験結果を示す．提案手法は，Bing が推薦できないマイナー意図のクエリを推薦することができ，マイナー検索の意図とクエリの曖昧さ回避の推薦に役立つことが証明されている．

5. まとめ

本論文では，マイナークエリの意図に対してもクエリ提案を生成する方法を提案した．我々の方法は，より一般的な用語を得るために元のクエリを拡張し，2 つのクエリ単語を接続できるフレーズを抽出する．関係フレーズを元のクエリに追加し，クエリ候補を生成する．実験結果は，我々のアプローチが有用性を示している．

将来の課題として，いくつかあげられる．

1. 提案手法は，二語のクエリに限定しているため，二語以上のクエリの推薦がまだ対応していない．将来では，二語以上のクエリを二語クエリの組み合わせで表現し，提案手法と同じような方法でクエリ推薦を行うことが考えられる．

2. この実験は．提案手法がいくつかのマイナーな検索意図のためにクエリを提案するのに有益であることを証明しているが，この実験では，一般的な 2 つの用語クエリすべてに対して有効であるということを証明していないため，今後はテストクエリを増やし，より完全な実験を行い，提案手法を評価する予定である．

文　献

[1] Ricardo A Baeza-Yates, Carlos A Hurtado, Marcelo Mendoza, et al. Query recommendation using query logs in search engines. In *EDBT workshops*, volume 3268, pages 588–596. Springer, 2004.

[2] Holger Bast and Ingmar Weber. The completesearch engine: Interactive, efficient, and towards ir & db integration. In *Third Biennial Conference on Innovative Data Systems*, pages 88–95, 2007.

[3] Sumit Bhatia, Debapriyo Majumdar, and Prasenjit Mitra. Query suggestions in the absence of query logs. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 795–804. ACM, 2011.

[4] Huanhuan Cao, Daxin Jiang, Jian Pei, Qi He, Zhen Liao, Enhong Chen, and Hang Li. Context-aware query suggestion by mining click-through and session data. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 875–883. ACM, 2008.

[5] Silviu Cucerzan and Ryen W White. Query suggestion based on user landing pages. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 875–876. ACM, 2007.

[6] Alan Feuer, Stefan Savev, and Javed A Aslam. Evaluation of phrasal query suggestions. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 841–848. ACM, 2007.

[7] Anil K Jain, M Narasimha Murty, and Patrick J Flynn. Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323, 1999.

[8] Bernard J Jansen, Amanda Spink, Judy Bateman, and Tefko Saracevic. Real life information retrieval: A study

of user queries on the web. In *ACM SIGIR Forum*, volume 32, pages 5–17. ACM, 1998.

[9] Rosie Jones, Benjamin Rey, Omid Madani, and Wiley Greiner. Generating query substitutions. In *Proceedings of the 15th international conference on World Wide Web*, pages 387–396. ACM, 2006.

[10] Yanan Li, Bin Wang, Sheng Xu, Peng Li, and Jintao Li. Querytrans: Finding similar queries based on query trace graph. In *Web Intelligence and Intelligent Agent Technologies, 2009. WI-IAT'09. IEEE/WIC/ACM International Joint Conferences on*, volume 1, pages 260–263. IEEE, 2009.

[11] Hao Ma, Haixuan Yang, Irwin King, and Michael R Lyu. Learning latent semantic relations from clickthrough data for query suggestion. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 709–718. ACM, 2008.

[12] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[13] Jonas Mueller and Aditya Thyagarajan. Siamese recurrent architectures for learning sentence similarity. 2016.

[14] Swarnadeep Saha, Harinder Pal, and Mausam. Bootstrapping for numerical open ie. In *ACL*, 2017.