クエリ中の単語の語義絞り込みによる動画検索精度の向上

平川 幸司[1]　菊池 康太郎[1]　植木 一也[2]　林 良彦[1]　小林 哲則[1]
[1] 早稲田大学理工学術院，[2] 明星大学情報学部

hirakawa@pcl.cs.waseda.ac.jp

1　はじめに

動画に対するアドホック検索においては，クエリが指定する事物が描写されているフレーム画像を適切に選別することが必要となる．一般にこのような処理は，事物を指定する単語の語義と対応付ける形であらかじめ構成した物体カテゴリの識別器により行われる．このため，クエリ中に含まれる事物を表す単語 (キーワード) を抽出し，その語義に応じて適切な識別器を選択することにより，最終的な検索精度の向上に寄与することが期待できる．

以上のような動機に基づき，本研究ではクエリ中のキーワードの語義曖昧性の解消･絞り込みを語義･概念の分散表現を用いることにより行う手法を数種類実装し，その精度が最終的な動画検索の精度にどのように影響するかを調べた．TRECVID AVS (Ad-hoc Video Search) タスクのデータセットを用いた実験によれば，複数のクエリに対する平均的な語義解消精度と最終的な動画検索の精度は中程度の相関を示した．また，複数の語義を許容する語義絞り込みにより，動画検索の精度向上が見られた．以上から，動画検索の精度向上のためにはクエリ中のキーワードの適度な語義絞り込みが有効である．

2　AVS タスク

動画アドホック検索に関するシェアードタスクである TRECVID AVS タスク[1]では，30 のクエリ (例: "Find shots of one or more people eating food at a table indoors") が与えられ，各クエリに対して，335,944 本の検索対象動画から適切な動画を検索することが求められる．参加システムによる検索結果はランキングにより表現され，正解との比較が行われる．比較においては，平均適合度 (mean AP: mean average precision) が評価尺度として用いられ，この結果によりタスク参加者の評価が行われる[2]．

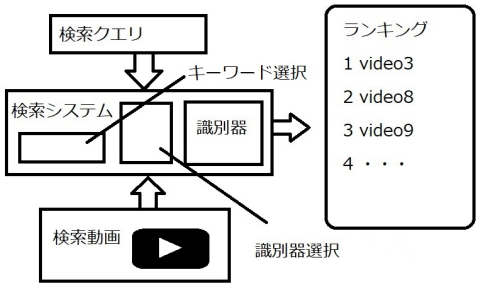

図 1: システムの概要

動画アドホック検索システムの一般的な構成を図 1 に示す．まず検索クエリからキーワードを抽出する．ここでキーワードとは，クエリに対して適切なフレーム画像に描写されていることが想定される事物を表す単語 (複合語含む) である．次に，各キーワードに対応する物体カテゴリ識別器（以下，識別器）を選択し，フレーム画像中に対象の事物が描写されているかを判定する．一般に識別器は，これを作成する際に用いたデータセットにおけるカテゴリ名によってラベル付けされており，多くのシステムにおいては，キーワードの字面と識別器のラベルの文字列的な一致がマッチングの基準として利用されている．このため，キーワードが多義をもつ場合，クエリにおける語義に適合しない識別器をも選択･利用してしまうという問題がある．

識別器の多くは，ImageNet[3]データベースに基づいて事前に構成される．ImageNet における画像は，WordNet における語義概念 (synset) ごとに整理されているため，WordNet を語義目録とするキーワードの語義曖昧性解消・絞り込みにより不適切な識別器の利用が阻止でき，最終的な動画検索精度の向上に寄与するものと期待できる．

[1] http://www-nlpir.nist.gov/projects/tv2016/tv2016.html

[2] 当研究室は 2017 年のタスクにおいて，自動走行で 2 位，手動走行で 1 位の成績をあげている [1].

[3] http://www.image-net.org/

3 語義曖昧性の解消

ある文脈に出現する単語の語義を定める処理は語義曖昧性解消 (Word Sense Disambiguation: WSD) [2] と呼ばれ，自然言語処理における中心的な課題の一つとなっている．WSD の手法は，教師付き学習による手法と学習によらない方法に大別できるが，前者の手法には対象タスクと適合した領域における語義タグ付きのコーパスが必要となる．このため，本研究では学習によらない方法を検討する．

学習によらない WSD は何らかの外部の資源を必要とする．その代表的な手法である Lesk アルゴリズム [2] は，ターゲットの単語が出現した文と語義の説明文・定義文を比較し，これらにおいてオーバラップする単語数の多い語義を選択する．

一方，統計的な自然言語処理においては，似た意味を持つ単語は出現文脈の分布が似ているという意味の分布仮説 (distributional hypothesis) [3] が大きな役割を果たしている．そこで，ターゲット単語のもつそれぞれの語義を仮定した場合の前後文脈との意味的類似性を用いることにより WSD を行う可能性を検討する．

このためには，単語の持つ語義，あるいは，それが指示する概念に対する表現が必要となる．本研究では，このために AutoExtend [4] と呼ばれる手法により構成された語義・概念の分散表現 [5] を利用する．この手法は，単語の分散表現と WordNet の辞書構造を利用して単語の分散表現と同一の空間に語義・概念に対する分散表現を導出するため，単語と語義，単語と概念といった異なる言語単位に対する意味的類似度を求めることが可能である．

なお，単語が複数の語義を有する場合，これらの語義の使用頻度には偏りがあることが知られている．この性質に基づき，最もよく用いられる語義を与えられた文脈にかかわらず選択するベースライン手法 (Most Frequent Sense: MFS[6]) が非常に強力であることが知られている．そこで本研究においても実験において比較する一手法として MFS を採用する．

4 提案する語義曖昧性解消手法

本研究では，語義曖昧性解消 (一つを選択)，語義絞り込み (複数語義を許容) の動画検索における効果について検討するが，まずは最適と考えられる語義を選択し，それを基準として複数の語義を残すかどうかという方略をとる．

このため，以下の 4 つの WSD 手法 (それぞれ DistLesk, DistSim, SimSum, SimMinMax と呼ぶ) を比較する．DistLesk 法以外の手法においては，ターゲット単語の語義・概念を表す表現が必要となるが，それぞれ AutoExtend 手法により導出された語義 (lexeme) ベクトル，概念 (synset) ベクトルを用いる．また，DistLesk 法では WordNet の語義に与えられている定義文を利用する．

4.1 Lesk アルゴリズムを拡張した手法:DistLesk

式 1 に示すように，クエリ中の各単語の分散表現ベクトル (以下，単語ベクトル) を平均することによって求めるクエリベクトル v^q とターゲットの単語の各語義 i に対する辞書中の定義文に含まれる単語ベクトルの平均によって求める．定義文ベクトル v_i^d の間のコサイン類似度 ($cosim$) を比較し，最大の類似度を与える語義 I を選択する．本手法は，単語の分散表現の利用による Lesk アルゴリズムの拡張とみることができる．実験では，クエリベクトルを求める際にターゲット単語を残す場合と除去する場合の双方を試した．

$$I = \arg\max_i cosim(v^q, v_i^d) \tag{1}$$

4.2 分布仮説を利用した手法:DistSim

文脈における周辺単語との類似度の分布の類似性を利用する本手法の概要を図 2 に示す．ターゲットの単語の表現として単語ベクトルを用いた場合の前後文脈の各単語の間との類似度を並べたベクトル (以下，文脈類似度ベクトル) s^w を考え，さらに，ターゲット単語の語義 i を仮定した場合の文脈類似度ベクトル s_l^i を考える．式 4 (文脈をターゲット語の前後 1 語ずつとしている) に示すように，s^w との間でもっとも高い類似度を示す s_l^i を与える語義 I を選択する．この手法では，(1) 単語ベクトルと語義・文脈ベクトルを直接に比較可能である，(2) ターゲット単語の単語ベクトルはこの単語の持つ語義ベクトルが加算されたものであるというという AutoExtend 法の特徴を利用している．実験においては，類似度分布を求める際に前後の単語数を 1 語ずつ～3 語ずつの場合を試した．

$$s^w = (cosim(v_{j-1}^w, v_j^w), cosim(v_{j+1}^w, v_j^w)) \tag{2}$$

$$s_l^i = (cosim(v_{j-1}^w, v_i^l), cosim(v_{j+1}^w, v_i^l)) \tag{3}$$

$$I = \arg\max_i cosim(s^w, s_i^l) \tag{4}$$

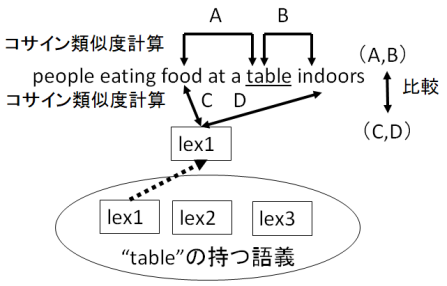

図 2: DistSim 手法の概要

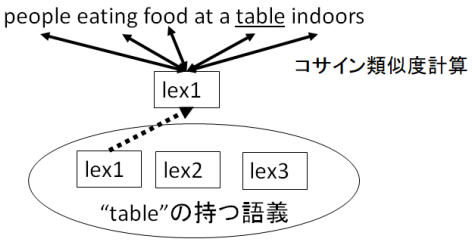

図 3: SimSum 手法の概要

4.3 文脈単語との類似度の合計値を利用する手法:SimSum

本手法 SimSum では式 5 に示すように，ターゲット単語の語義 i を仮定したときのクエリ中の各単語の単語ベクトルとのコサイン類似度の合計を計算し，最も高い類似度を与える語義 I を選択する．本手法には，分散表現には前後文脈の単語を予想するのに有用な含まれる傾向があるという性質を利用している．

$$I = \arg\max_i \sum_j cosim(v_i^l, v_j^w) \tag{5}$$

4.4 文脈単語との類似度の最小値を利用する手法:SimMinMax

本手法は上記の SimSum 手法とほぼ同様であるが，式 6 に示すように合計ではなく最小値を求め，これが最大となる語義 I を選択する．

$$I = \arg\max_i \min\{cosim(v_j^l, v_j^w)\} \tag{6}$$

5 実験

5.1 実験設定

TRECVID2016 AVS タスクのデータを用いて，クエリ中のキーワードの語義選択を行わない (候補識別器を全て利用する) 場合，理想的な語義選択が行われた場合，MFS による手法，本研究で提案する 4 種類の手法による語義曖昧性精度，動画検索精度を比較した．キーワード抽出では NLTK の tokenizer と stopword list を利用した．

キーワードの正解語義付与: 30 件のクエリに含まれる延べ 73 箇所のキーワード出現 (異なり数:53) に対して，手動により WordNet における正解語義を付与した．ただし，複数の語義が適切な場合はこれらを残した．また，不必要なキーワードには正解語義を付与しなかった．なお，候補となる語義が一つしかないキーワード出現は 41 箇所であった．

動画検索 [1]: 識別器は対象動画それぞれに対して評価値を持ち，一つのキーワードに複数の識別器が選択された場合はその値を平均化する．各クエリごとに全ての動画に対して以下のような計算を行い動画のスコアを求め，スコアの降順にランキングを作成する．ここで f はその動画のスコア，n はクエリ中のキーワード数，x_i は識別器の持つ評価値，t_i は MSCOCO データセット[4]から求めたキーワードの idf である．

$$f = \prod_{i=0}^{n} x_i^{t_i} \tag{7}$$

5.2 評価指標

評価指標として語義曖昧性解消の精度を評価するための解消成功率と動画検索の精度を評価するための平均適合率を用いた．

解消成功率 R: 人手で付与した正解語義と比較し，解消成功率 R を式 8 で計算する．ここで，x は 30 クエリに含まれるキーワードの数で，y はキーワードの内不適な語義を選択したキーワードの数である．

$$R = \frac{x - y}{x} \tag{8}$$

平均適合率 mAP: 動画検索により得られたランキングと正解を比較して Average Precision (AP) を算出し，これを全てのクエリで平均する[5]．

5.3 主要な実験結果と考察

WSD 手法・精度と動画検索精度を比較する実験結果を表 1 に示す．今回提案した手法の結果の中に，解消成功率 R が高いが，mAP が低いものや，その逆で，

[4] http://cocodataset.org/

[5] 参考までに，TRECVID AVS 2017 における自動タスクの最高精度 (アムステルダム大学) は 0.2065 であった．

表 1: 実験結果（*付きの手法は識別器を複数選択する場合がある）

手法名			R	mAP
ベース*			0.6301	0.17792
手動*			1.0000	0.19218
MFS			0.8493	0.17478
DistSim	1 単語	Lexeme	0.7260	0.17087
		Synset	0.7534	0.17106
	2 単語	Lexeme	0.7945	0.17348
		Synset	0.7945	0.15812
	3 単語	Lexeme	0.8082	0.17438
		Synset	**0.8219**	0.16529
SimSum	Lexeme		0.8082	**0.17893**
	Synset		0.7808	0.17851
SimMinMax	Lexeme		0.8082	0.17787
	Synset		0.7671	0.17788
DistLesk	KW あり		0.7397	0.15385
	KW なし		0.7534	0.15051

表 2: 語義絞り込みによる動画検索精度 (mAP)

手法名		(表 1 再掲)	評価値	類似度
DistSim	Lexeme	0.1744	0.1767	0.1762
	Synset	0.1653	0.1695	0.1805
SimSum	Lexeme	0.1789	0.1789	0.1816
	Synset	0.1785	0.1784	0.1785

R が低いにもかかわらず，mAP が高いものがある．これは，語義曖昧性解消に成功したキーワードが手法により異なり，結果として使用された識別器が異なることによる．しかし総合的には WSD は動画検索精度の向上に有効であり，両者の間には中程度の相関関係（相関係数:0.428）が認められた．

WSD の精度は分布仮説を利用した DistSim 手法の精度が良好であったが，従来の WSD 研究における知見と同様に MFS を超えることはできなかった．また，類似度分布を作成するのに使用した文脈単語数を 4 単語以上に増やしても精度の向上は見られなかった．

5.4 語義絞り込みによる検索精度向上

言語的な WordNet の語義区分は，視覚的に区別すべきカテゴリと必ずしも一致しない．WordNet の語義の粒度は細かすぎるという批判もあるため，複数の語義を許容し，複数の識別器を利用することにより，動画検索の精度が改善される可能性がある．本研究では語義曖昧性解消により選択された語義と類似する語義を残すこととし，WSD における評価値の差，語義・概念ベクトルにおける類似度という 2 つの観点を比較した．語義解消成功率 R が最も良かった DistSim 手法 (前後の単語数:3) と mAP の最大値を与えた SimSum 手法において複数語義を許容し，mAP を算出した．評価値を利用した絞り込みでは選択語義の評価値の 99% を閾値とし，類似度を利用した絞り込みでは，DistSim 手法:0.5，SimSum 手法:0.7 と設定した．

表 2 に結果を示すように，動画検索精度の向上が確認できた．すなわち，本研究のタスクである動画検索においては，語義は解消するのではなく，適度に絞り込むことが効果的である．

6 むすび

検索クエリ中のキーワードの語義曖昧性解消・絞り込みによる動画検索精度の向上を試み，学習によらない WSD においては語義・概念の分散表現の利用が有効であること，検索精度の向上のためには，適切に複数語義を許容する語義絞り込みが効果的であることを確認した．前段のキーワード選択の段階で動画検索において不要なものを削除すること，場合によってはクエリ内に現れていない関連キーワードを追加する，などの改善を行うことにより，手動による理想的な結果に迫る精度が達成できるものと考えられる．

謝辞

本研究は JSPS 科研費 (17H01831) の助成を受けた．

参考文献

[1] K. Ueki, K. Hirakawa, et al., ”Waseda Meisei at TRECVID 2017:Ad-hoc video search”. TRECVID 2017 (2017).

[2] R. Navigli. ”Word sense disambiguation: A survey”. ACM Comput. Surv. 41, 2, 169.

[3] J, R. Firth, ”A synopsis of linguistic theory 1930-1955”. Studies in Linguistic Analysis. Oxford: Philological Society: 132. (1957).

[4] S. Rothe and H. Schütze. ”AutoExtend: Extending word embeddings to embeddings for synsets and lexemes”. *Proc. of ACL2015*, pp.1793-1803, (2015).

[5] 金田 健太郎，小林 哲則，林 良彦．”単語，語義，概念: 意味タスクにおける分散表現の適用性”．2017 年度人工知能学会全国大会．

[6] E. Agirre, and P. Edmonds. *Word Sense Disambiguation: Algorithms and Applications*. Springer. (2006).