2005－SLP－57（24）
2005／7／16

ベクトルの非類似度を用いて 複数表現の接続詞を自動決定するお天気情報システム

飯田朱美、相川清明

ake@media.teu.ac.jp, aik@media.teu.ac.jp

〒192-0982 八王子市片倉町 1404-1
東京工科大学メディア学部

概要

我々は、ベクトル空間法を用いて、計測した多変量な情報を話し手聞き手の双方にとって理解しやすい表現に置き換えて出力する手法を提案した。そして、評価システムとして、温度や湿度などの気象情報を「さわやか」「うっとおしい」などの、日常我々が気象情報を他者に伝える時に使う感覚的に理解しやすい表現に置き換えて出力するシステム、Weather Reporter を実装した。本システムは、二つの表現を接続して出力することができるが、これまでは全ての組合せを順接接続詞、「そして」で接続していたため、共起するとは考えにくい組合せの排除や語義に応じた接続詞の使い分けが課題となっていた。そこで、実際にはどのような接続形態が在り得るのかをアンケート調査を行って調査し、二つの表現のテンプレートベクトル間の非類似度を求め、複数の候補から接続詞を自動決定する手法を考案し、実装し、評価したので報告する。

VECTOR-BASED WEATHER REPORTING SYSTEMS USING DISSIMILARITY MEASURING ALGORITHM TO AUTOMATICALLY ASSIGN CONJUNCTIVES

Akemi Iida, Kiyoaki Aikawa

School of Media Science
Tokyo University of Technology

Abstract

In previous research, we proposed an approach for describing multivariate environmental information as it might be expressed by humans using easy to understand day-to-day phrases. In its prototype, the Weather Reporter took meteorological data as input and produces expressive phrases such as “refreshing” or “muggy” as output. This system could also conjoin two phrases such as “hot and muggy”, but the first version of the reporter could only assign the conjunction ‘and’ regardless of the meaning of the two phrases. Hence, we conducted a survey to find out how people typically conjoin two meteorological phrases. This paper reports on the findings from the survey and also proposes a new dissimilarity measuring algorithm that automatically assigns the conjunction additive ‘and’ or adversative ‘but’ to conjoin two sentences depending on the meanings of the two phrases that have been selected by the system to be combined.

1. はじめに

我々は、ベクトル空間法を用いて、計測した多変量な情報を話し手聞き手双方にとって理解しやすい表現に置き換えて出力する手法を提案した。そして、評価システムとして、温度や湿度などの気象情報を「さわやか」「うっとおしい」などの、日常我々が気象情報を他者に伝える時に使う感覚的に理解しやすい表現に置き換えて出力するシステム、Weather Reporter を実装し、報告してきた[1][2]。本システムは、二つの表現を接続して出力することができるが、これまでは全ての組合せを順接接続詞、「そして」で接続していたため、共起するとは考えにくい組合せの排除や語義に応じた接続詞の使い分けが課題となっていた。そこで、実際にはどのような接続形態が在り得るのかをアンケート調査を行って調査した。本稿では、調査結果を報告するとともに、二つの表現のテンプレートベクトル間の非類似度を求めて、複数の候補から接続詞を自動決定する方法を提案する。

2. ベクトル空間法による気象表現の選択

本手法では、入力される温度、湿度などの環境情報と、内部的に保持される環境表現語句をともに 62 次元のベクトルとして表現した。本稿では前者を入力ベクトル、後者をテンプレートベクトルと呼ぶ。環境表現語句については文献[3][4]、および、Web ページから 149 の語や句を収集した。本稿ではこれらを基本エントリーと呼ぶ。

ベクトルの次元は 62 次元で、その内訳を表 1 に示す。テンプレートベクトルの適合値は、表 1 の各環境情報項目に該当すれば 1 、該当しなければ 0 を基本としたが、ある特定のベクトル要素に重みを置く場合、1 より大きい値も用いた。環境情報項目には、月、日、時間、天気のようにカテゴリカルなベクトル要素と、気温、湿度、風力といった順序的なベクトル要素が混在する。後者の場合には変化域をいくつかの区間に分け、複数のベクトル要素を割り当てた。例えば、気温には、-4° C から 38° C まで 2 度刻みで 22 個のベクトル要素を割り当てた。そして、同じ環境情報の複数の区間に適合する場合、重みを変えてベクトル要素の値を設定した。例えば「蒸し暑い」の場合には 28, 30, 32, 34(° C)にあたるベクトル要素の値を 1.8, 2.0, 2.0, 1.8 とした。

次に、外部から取り込んだ環境情報も同様にベクトル化し、この入力ベクトルと複数のテンプレートベクトルの余弦を類似度として求め、最も類似度の高いベクトルに対応する表現を選択した。

表 1. 環境情報のベクトルの要素

環境情報		環境情報項目	要素
月	C	1 月から 12 月	12
日	C	初旬、中旬、下旬	3
時間	C	午前、午後、夜	3
天気	C	快晴、晴れ、曇り、小雨、雨、雪、雷	7
気温	O	-4° C から 38° C まで 2° C 刻み	22
湿度	O	0 から 100% まで 10%刻み	11
風力	O	無風、微風、弱風、強風	4
合計			62
C：カテゴリカル、O：順序的ベクトル			

3. 合成ベクトルの生成方法

一つの表現を用いるよりも、二つ以上の表現を併用すると、より適切な表現ができる場合もある。例えば「べたつく」と「かんかん照り」を接続させることにより、「ものすごく気温も湿度も高い」ということを表現できる。そこで、Weather Reporter では、複数のテンプレートベクトルを合成し、選択候補に加えた。原理的には基本エントリーの数までのベクトルの合成が可能であるが、二つのベクトルの合成までを実装している。

ベクトル空間においては、二つの基本エントリーのテンプレートベクトル a, b の合成ベクトルは a,b 間に挟まれた領域に示されると考えることができる。合成ベクトルの算出方法は、二つの基本エントリーのテンプレートベクトルの和としてではなく、論理和として求めた。また、片方または両方の要素の値が 1 でない場合は 1 とした。全ての基本エントリー(N)を対象に合成ベクトルを作成し、その数は N*(N-1)/2 個となった。

4. 色々な合成ベクトル

これまでの実装では、全ての合成ベクトルにおいて、「べたつく、そして、かんかん照りだ」のように、順接接続詞「そして」を用いて、二つの基本エントリーを接続していた。しかし、中には、これではうまく行かないものもある。例えば、「暑い、そして、寒い」とは、顔に熱風を当てられ、かたや足は凍り水につけられるなどしない限り、普通の状態では言わないだろ

う。そうかと言って、「暑いが寒い」とも同様な環境に強制的に置かれない限り言わないだろう。では、「暑い」と「過ごしやすい」ではどうだろうか。温度が高いわりに湿度が低い時には不快指数も低くなるため、意味的に相反していても言えそうな組合せである。このように考えると、組合せの中には意味的に相反する組合せや比較的意味が近い組合せがあることがわかる。システムのふるまいを矯正するために、いくつかの組合せについて日常我々がどのように発話しているかを考えていくうちに、組合せのパタンは、以下のように分類することができそうだとわかってきた。

[Type 1] 共起するとは考えにくい対義的組合せ（「暑い」と「寒い」）→接続させない。

[Type 2] 共起し得る対義的組合せ（「暑い」と「過ごしやすい」→逆接接続詞で接続。

[Type3] 同義的（「すがすがしい」と「気持ちいい」）→順接接続詞で接続。

5. アンケート調査

4 節で述べた分類は著者らの主観的分類である可能性があり、また、一般的な分類であっても、接続形態は人により異なる可能性があるため、人々が日常どのように使っているかをアンケート調査することにした。

5.1. アンケート 1

このアンケートでは 5 選択肢から 1 肢選択するよう依頼した。

5.1.1. 回答者

日本語を母語とする 22 人（男 15、女 7 人）。年代：10 代後半 3 人、20 代 6 人、30 代 3 人、40 代 5 人、50 代 1 人、60 代 2 人、70 代 2 人。

5.1.2. 質問の形式

以下の形式で質問した。

「仮に、あなたが Sa と Sb という表現しか使えなかったとします。「○月、温度○度、湿度○%、晴曇雨のいずれか、風力○○ぐらい」のように、Sa と Sb の中間のような気候の時、この二つの表現を使って、どう言いますか？」

5.1.3. 提示した表現と環境条件（提示順）

(1) [Type 1]「暑い」と「寒い」
5 月、温度 22℃、湿度 50%、快晴、微風

(2) [Type 3]「暖かい」と「さわやかだ」
4 月、温度 20 度、湿度 40%、快晴、微風

(3) [Type 2]「暑い」と「過ごしやすい」
7 月、温度 28 度、湿度 50%、快晴、微風

(4) [Type 3]「寒い」と「北風が吹いている」
1 月、温度 0 度、湿度 10%、曇り、強風

(5) [Type 3]「暑い」と「むしむしする」
7 月、温度 30 度、湿度 90%、快晴、無風

(6) [Type 2]「からりと晴れている」と「身を切るような寒さ」
2 月、温度-2 度、湿度 10%、快晴、無風

(7) [Type 3]「すがすがしい」と「気持ちいい天気」
5 月、温度 24 度、湿度 30%、快晴、微風

5.1.4. 選択肢

選択肢には、(1)から(7)それぞれの提示表現を用いて、以下の選択肢を用意した。これらを図式化すると、図 1 のようになる。

(1) Sa （例：「暑い」）
(2) Sb （例：「寒い」）
(3) Sa and Sb （例：「暑いし、寒い」）
(4) Sa but Sb （例：「暑いが、寒い」）
(5) Sa nor Sb （例：「暑くも寒くもない」）

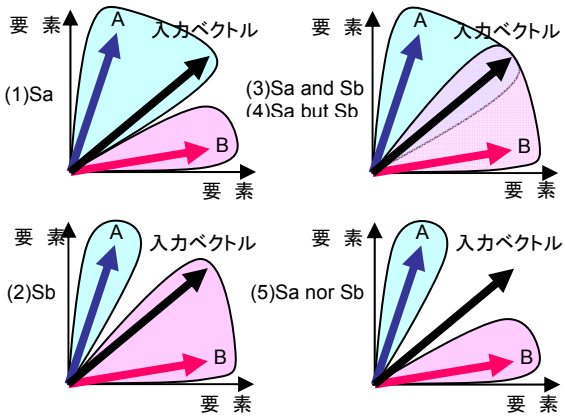

図 1. 選択肢のパタン

表 2. アンケート 1 の結果

		Sa	Sb	Sa and Sb	Sa but Sb	Sa nor Sb
Type 1	「暑い」「寒い」	14%	0%	9%	0%	77%
Type 2	「暑い」「過ごしやすい」	27%	14%	5%	50%	5%
Type 2	「からりと晴れている」 「身を切るような寒さ」	9%	18%	14%	59%	0%
Type 3	「暖かい」「さわやかだ」	18%	18%	59%	5%	0%
Type 3	「寒い」「北風が吹いている」	36%	5%	59%	0%	0%
Type 3	「暑い」「むしむしする」	9%	14%	77%	0%	0%
Type 3	「すがすがしい」「気持ちいい天気」	5%	45%	50%	0%	0%

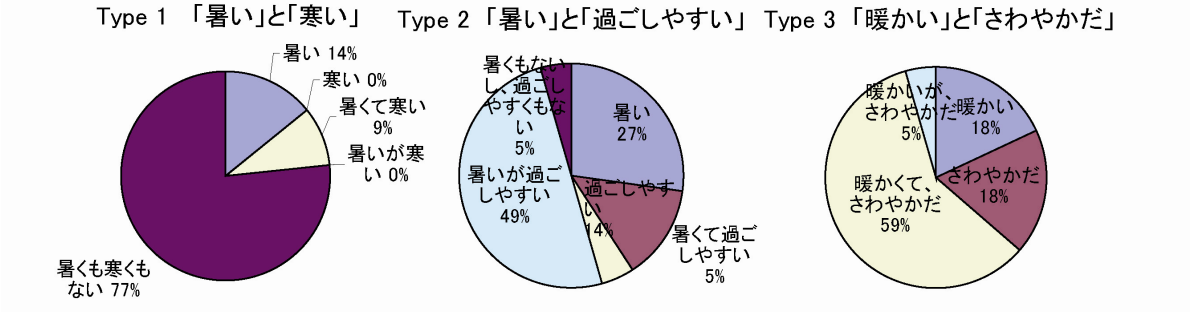

図 2. タイプ別回答結果例

表 3. アンケート 2 の結果

	Sa	Sb	Sa and Sb	Sa but Sb	Sa nor Sb
Type 1	30%	0%	10%	40%	100%
Type 2	70%	70%	10%	100%	0%
Type 3	98%	93%	93%	0%	0%

5.1.5. 回答

回答結果を表 2、および、図 2 に示す。Type 1 の組合せでは Sa nor Sb の「暑くも寒くもない」を選択した回答者が 77%と圧倒的に多かった。Type 2 では、「暑いが、過ごしやすい」が、50%、「からりと晴れるが、身を切るような寒さ」を選んだ回答者が 56%と、Sa but Sb を他の 4 選択肢を大きく上回った。Type 3 では、提示順に 59%, 59%, 70%, 50%が、Sa and Sb を選択した。

5.2. アンケート 2

アンケート 1 と同じ内容の設問に対して、使う表現には○、使わない表現には×をつけてもらった。

5.2.1. 回答者

日本語を母語とする 10 人（男 6 人、女 4 人）。年代：10 代後半 3 人、20 代 6 人、40 代 1 人。

5.2.2. 回答

回答結果を見ると、Type3 の同義的組合せの場合は、その組合せを構成している表現のどちらかでも表現でき、両者を順接接続詞で組合せても表現できることが示されている。すなわち、Sa, Sb, Sa and Sb のいずれでも表現できることが示されている。Type2 では、構成語のどちらかでの表現よりも逆接接続詞で接続した表現の方が使われることが示された。Type1 については、Sa nor Sb が 100%、Sa but Sb が 40%で、構成語

のいずれかでの表現を使うという回答は他の２タイプより少なかった。

5.3. 考察と方針

アンケート結果は、実験前の仮定と一致していた。そこで、Type 2 は逆接接続詞、Type 3 は順接接続詞で接続して出力する必要がある。Type 1 については、このような組合せは排除するアルゴリズムを検討する必要がある。

6. 非類似度を用いた接続詞の自動決定

アンケート結果により、合成ベクトルを接続する場合には接続したい基本エントリーの語義が同義か対義かによって、接続詞を逆接と順接の二通り用意しておく必要性が確認できた。また、共起するとは考えにくい対義の組合せの場合は選択候補として合成しないなどの処理も必要だということがわかった。今回は語義によって、順接と逆接の接続詞を自動的に決定する手法を考え実装した。

6.1. コンセプト

これまでの Weather Reporter は二つの表現をそれらの語義に関係なく、全て、順接接続詞「そして」で接続した。しかしながら、前節で述べたように、対義関係にある２つの表現の場合には、逆接接続詞で接続したい。しかし、全ての組合せに対して、二つの基本エントリーの関係が同義なのか対義なのか記述していくことは、組合せの数が膨大なだけに困難であり、現実的ではない。そこで、本研究では、二つの基本エントリーのテンプレートベクトルの各環境情報（表１参照）の非類似度を求めることで接続詞の自動決定を可能にした。

6.2. 非類似度の自動決定手法

それぞれの環境情報を「部分空間」とみなし、それぞれの部分空間におけるベクトルを「環境要素ベクトル」と呼ぶことにする。表１において、月、日、時間、天気などのカテゴリカルなベクトルは二つの基本エントリーのベクトル要素の正弦を計算することで求めることができる。一方、気温、湿度、風力などの順序的なベクトルは、２節で述べたように、いくつかの連続した項目の要素が０以上の値を持つために、非類似度はそれぞれの環境要素ベクトルを統計的分布とみなし、それぞれの分布間の距離を計算することで、求めることができる。二つの分布間の距離は分布パタンの相互偏差として求めた。実際には、この値は二つの環境要素ベクトルの相互相関の平均として求めた。従って、k 番目の環境要素ベクトル u_k と v_k は式(1)で求められる。

$$d_k = \left| \frac{\sum_{i=-L_k}^{L_k} i c_k(i)}{\sum_{i=-L_k}^{L_k} c_k(i)} \right| \qquad (1)$$

$$c_k(i) = \frac{1}{L_k} \sum_{j=-L_k}^{L_k} u_k(j) v_k(j+i),$$

式(1)において、c_k は環境要素ベクトル u_k と v_k の相互相関関数であり、L_k は環境要素ベクトルの長さである。iu_k と v_k が全く同じであれば、非類似度 d_k は０になり、二つの環境要素ベクトルは同じということになる。この式は相互相関関数を一種の確率密度関数とみなしたときの分布の期待値であり、順序的ベクトルの位置ずれの期待値となっている。

その後、全ての環境要素ベクトルの非類似度の平均を出し、その値が、０に近ければ、Weather Reporter は二つの基本ベクトルは対義ではないと判定する。一方、その値が 1 に近ければ、システムは両ベクトルは同義と判定する。現在のところ、カテゴリカルなベクトルは非類似度計算から除外し、閾値はヒューリスティックに 0.4 と決定した。接続は順接の場合は一つ目の基本エントリーに「し」を付け、逆接の場合は「が」をつけるようにした。

6.3. 評価

以前と同様の環境（Windows Professional XP, MATLAB 7.1）のもとで、Weather Reporter に 6.2 で述べた非類似度決定部を MATLAB により実装し、評価を行った。まず、アンケートで提示した 7 つの組合せ中、Type 1 の「暑い」と「寒い」の組合せを除く 6 つについて、アンケート結果と同じ結果が出るかを調べた。他の基本エントリーの影響を避けるために、対象となる基本エントリーとそのテンプレートベクトルのみを保持させて、評価を行った。そして、GUI からアンケートで提示した気象条件を入力した。(4)の「暑い」と「すがすがしい」の組合せでは、合成ベクトル自体が選択されず、「暑い」が選択されたが、その他は期待通りの結果を出力した。

(1) Type 1 につき検証対象から除外。
(2) 暖かいし、さわやかだ　0.0923

表 4　順接と逆接接続詞の自動決定評価

入力された気象条件	非類似度	出力表現	結果
1 月, 上旬, 朝, 快晴, -4℃, 湿度 0%、無風	0.478	身を切る寒さだが、のどかだ	OK
5 月, 上旬, 朝, 快晴, 22℃, 湿度 0%、無風	0.0962	すがすがしいし、気持ちいい天気だ	OK
5 月, 上旬, 昼, 快晴, 16℃, 湿度 30%、無風	0.3679	のどかだが、青嵐が吹いている	OK
11 月, 下旬, 昼, 小雨, 20℃, 湿度 50%、弱風	0.1966	冷ややかな風が吹くし、木枯らし一号が吹くみたい	OK
1 月, 中旬, 朝, 曇り, 6℃, 湿度 20%、弱風	0.2849	寒さが加わるし、北風が吹いている	OK
2 月, 下旬, 昼, 曇り, 6℃, 湿度 60%、強風	0.4011	光の春だが、北風が吹いている	OK
3 月, 中旬, 朝, 曇り, 16℃, 湿度 50%、強風	0.0621	肌寒いし、うそ寒い	OK
4 月, 中旬, 朝, 晴れ, 20℃, 湿度 50%、微風	0.0234	すがすがしいし、ニシン曇りだ	NG
7 月, 中旬, 朝, 雷, 32℃, 湿度 80%、弱風	0.2683	雨がザーッと降るし、嵐のような天気だ	OK
7 月, 中旬, 夜, 快晴, 32℃, 湿度 80%、無風	0.1081	べたつくし寝苦しい	OK

（気象条件は左から月、日、時間、天気、気温、湿度、風力）

(3) 暑い（合成ベクトルは選択されず）
(4) 寒いし、北風が吹いている　0.2843
(5) 暑いし、むしむしする 0.1390
(6) からりと晴れているが、身を切るような寒さだ　0.5039
(7) すがすがしいし、気持ちいい天気だ 0.0889

さらに 149 個の基本エントリー全てと、それらのテンプレートベクトルを保持させて、10 回の試行を行ったところ、良好な結果を得ることができた。この結果を表 4 に示す。

7. 結論

我々は、ベクトル空間法を用いて、計測した多変量な情報を話し手聞き手双方にとって理解しやすい表現に置き換えて出力する手法を提案し、評価システムとして、温度や湿度などの気象情報を「さわやか」「うっとおしい」などの、日常我々が気象情報を他者に伝える時に使う感覚的に理解しやすい表現に置き換えて出力するシステム、Weather Reporter を実装した。このシステムでは二つ以上の表現を組み合わせて出力することができるが、これまでは全ての組合せを順接接続詞、「そして」で接続していたため、共起するとは考えにくい組合せの排除や語義に応じた接続詞の使い分けが課題となっていた。そこで、本稿では、人の日常の使用状況調査をもとに、両者の非類似度を用いて、複数の候補から接続詞を自動決定する方法を提案した。今後は共起するとは考えにくい組合せの自動排除の方法を検討する予定である。

8. 謝辞

日頃、ご助言いただきますメディア学部飯田仁教授に感謝致します。また、アンケートにご協力頂いた皆様に感謝致します。

9. 参考

[1] 飯田朱美, 上野嘉人, 松浦良平, 相川清明, 「ベクトル空間法を用いたイメージを想起させるお天気表現システム」, 情報処理学会 第 109 回ヒューマンインタフェース・第 52 回音声言語情報処理共催研究会 HIS109SLP52, pp. 113-18.
[2] Iida, A., Ueno, Y., Matsuura, R., Aikawa, K. “A Vector-based Method for Efficiently Representing Multivariate Environmental Information”, *In proceedings of ICSLP 2004, Cheju, Korea, pp.269-272 Proc., 2004.*
[3] 日本放送協会編, 『最新気象用語ハンドブック』, 日本放送協会(NHK ブックス), 1986.
[4] 気象庁ホームページ (n.d.). from http://www.jma.go.jp/JMA_HP /jma/index. html