動画中の人の動作を入力情報とする動的計画法を用いた言語生成モデル

小林 瑞季[†, a]　小林 一郎[†, b]　Sergio Gudarrama[††, c]　麻生 英樹[†††, d]

† お茶の水女子大学大学院 人間文化創成科学研究科 理学専攻 情報科学コース

††Electrical Engineering and Computer Sciences, UC Berkeley

††† 産業技術総合研究所 知能システム研究部門

a) *kobayashi.mizuki@is.ocha.ac.jp*　b) *koba@is.ocha.ac.jp*　c) *sguada@eecs.berkeley.edu*　d) *h.asoh@aist.go.jp*

概要　本研究では，視覚情報からそれを説明するテキストを確率的に生成する手法を提案する．視覚情報として，Kinect カメラによって捉えられた人の動きの時系列データを採用した．得られた時系列データは，数段階の次元圧縮手法を経たのち，機械学習に適した形に成形される．その後，処理された時系列データとそのデータの示す動作の中間表現のペアに対して，線形対数モデルを用いた機械学習を行う．テキスト生成のための言語資源としては，人の動作のさまざまな言語表現を収集し，それぞれの動作に対して構築されたバイグラムモデルを使用する．本手法では，観測された時系列データから中間表現を選択し，選択された中間表現に対応したバイグラムモデルを選択し，さらに選択されたバイグラムモデルに動的計画法を適用することでテキストを生成する．

キーワード　Kinect, 時系列データ，SAX，対数線形モデル，バイグラムモデル，動的計画法

1　はじめに

近年，大量の動画像データを取得することが容易になってきている．一方で，大量に収集したデータを有効に利活用出来ているとは言えない．例えば，監視カメラの動画像データに映る内容を把握するためには，全てを人目で見る必要があるが，データの多さに応じた時間を要してしまう．もし，大量の動画像データから特徴的なイベントを捉え，またそのイベントを言葉として表現することが出来たら，動画像データに映る内容を簡単に把握できるとともに，言葉で動画像中のイベントの検索も行うことができると考える．そこで本研究では，動画像の情報を入力とした確率的なテキスト生成手法を提案する．

2　関連研究

マルチメディア情報を入力としてテキスト生成を行う関連研究として，Ding ら [1, 2] はインターネット上のビデオクリップを説明するテキストの要約生成システムを提案している．また Tan ら [3] の研究では，ビデオコンテンツの視聴覚情報をクラスタリングし，その結果をテンプレートに当てはめることで，ビデオの説明文を生成する手法を提案している．小林ら [4] は部屋の中にいる人の振る舞いを言語化する手法を提案している．さらに，Barbu ら [5] はショートビデオの説明文を生成するシステムを提案している．これらの説明文は，誰が何に何をしたのか，どこでどうやって行ったのかを説明している．またこのシステムのテキスト生成の手法には，簡単な文法を加味したテンプレートベースのテキスト生成の手法が採用されている．

柔軟なテキスト生成に関しては，確率的なアプローチでテキスト生成を行う研究が数多くされている．Lapata[6] は領域固有のテキストコーパスから順序の制約を学習するモデルを構築することで，いくつかある候補の中でもっともよい並びを生み出すアルゴリズムを提案している．Belz と Kow ら [7, 8] は生成空間の統括的なモデルを用いた確率的生成手法を用い，天気予報のテキストを生成できるシステムを提案している．また，Lu ら [9] は，意味表現と自然言語の両方を木構造にエンコードしている混合木によるテキスト生成モデルを提案し，最新の自然言語生成モデルよりも良い結果を得られている．

また，本研究に関連の深い研究として，Liang ら [10] や Angeli ら [11]，Konstas ら [12, 13] の研究が挙げられる．Liang ら [10] は，テキストと意味との関係を学習する手法を提案しており，そこでは，イベントはデータベースのレコードで表せると仮定し，レコードと自然言語で表記された説明文との関連を機械学習によって取得する．Angeli ら [11] は， Liang ら [10] が提案したモデルに基づく潜在情報と表層情報をテキスト生成する手法を提案している．また，Konstas ら [12, 13] は，入力情報固有の構造を説明する確率的な自由文脈文法を定義しており，Liang ら [10] や Angeli ら [11] と同様に，データベースのレコードと説明文を用いている．彼らは，重みを加えたグラフによって文法を表現し，また与えられた入力に対しもっとも適切な導出木を見つけることでテキスト生成を行う

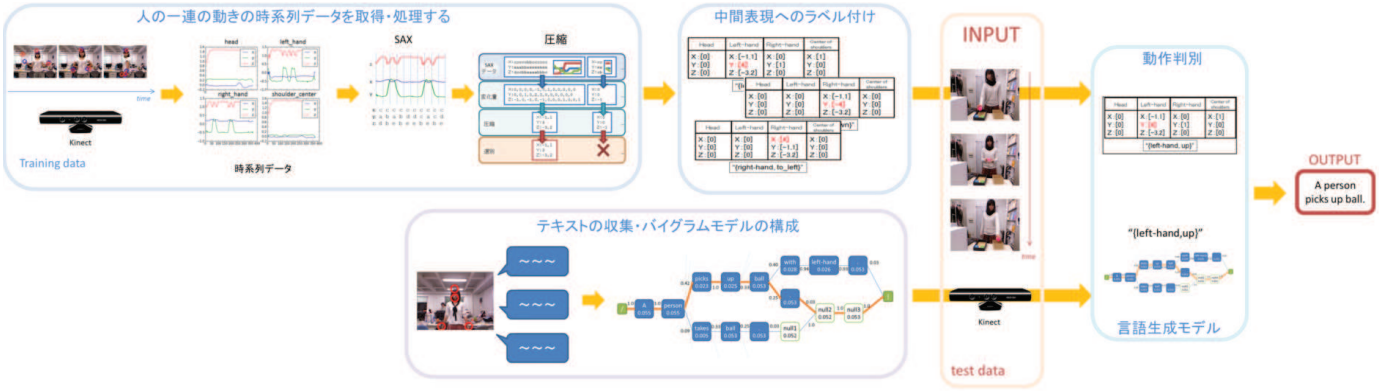

図 1 動画像を入力とする確率的テキスト生成の枠組み

本研究では，これらの関連の高い研究の考え方を参考にし，時系列データと動作を表す自然言語の説明文との対応を学習させるために対数線形モデルを採用した．提案するテキスト生成の手法は単純であり，生成に文法を必要とするような複雑な文は生成することができないが，動画像を入力とし，単純ではあるが一般的に使用される表現で構成される文を容易に生成することが可能である．

3 研究概要

本研究の概要を図 1 に示す．まず，Kinect[1] がもつ人の骨格を追跡するライブラリを用いることで，人の動きを時系列データとして取得する．取得された時系列データはいくつかの次元圧縮作業を行い，データと自然言語の仲立ちをする中間表現とともにデータベースに格納される． その後，データベース内に蓄積された時系列データと中間表現の対応関係を機械学習することで，動作判別器を生成する．テキスト生成に用いられる言語資源は，人の動作の表現を被験者実験によって収集し，それぞれの中間表現に対してバイグラムモデルを構築する．これにより中間表現を選択すると，その中間表現に対応したバイグラムモデルが選択され，そのモデルに動的計画法を適用することで，人の動作を表現するもっともらしい語の組み合わせを選ぶことができる．

3.1 時系列データ処理

人間の動作の時系列データは，Kinect カメラを用いて取得する．Kinect の開発元である MicroSoft 社は，人間の骨格を推定できる標準ライブラリも提供しており，そのライブラリを用いると人の関節の 3 次元位置情報を推定することができる．本研究では，RGB 画像と深度センサー，またそれらを用いた人物の関節位置推定も用い，RGB 動画像と人物の頭・肩の中心・右手・左手の 4 箇所の xyz 座標の時系列データを取得する (図 2 参照)．

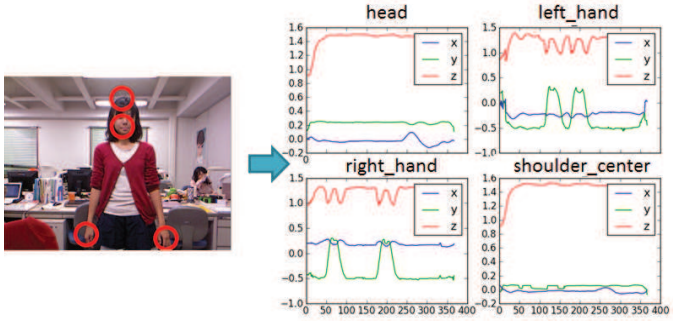

図 2 Kinect を用いた時系列データ取得

人の骨格を追跡することで得られた時系列データは，Symbolic Aggregation approXimation（SAX）[Lin 2003] を使い，文字列に変換する．

SAX によって変換して得られた文字列から動作とみられる個所を取り出す．ここでは，ある動画像データ中の全ての文字列において一つ前の文字から変化がなければ「動きがない」，変化があれば「動きがある」とみなす (図 3 参照)．

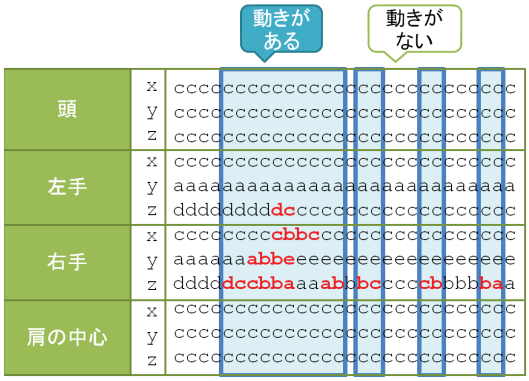

図 3 動きの抽出例

[1] http://www.microsoft.com/en-us/kinectforwindows/

その後,「動きがある」とみなされた個所の文字列を変化量 (図 4 中のアルファベットの下の数値) に変換し,圧縮する (図 5 参照).これは同じ動作でも位置やスピードによっては文字列がある一定の間隔でずれたり文字列の長さが変化したりしてしまい,同じ動きとして学習されないためである.これにより,一定の間隔でずれてしまったものも長さが違うものでも,同じ動きとしてとらえることを可能とする.また,より特徴的な動作を抽出するために,圧縮された変化量うち最大の大きさが 2 未満を示す動きは取り除く (図 5 参照).

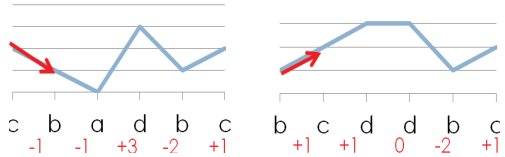

図 4 文字列の変化量

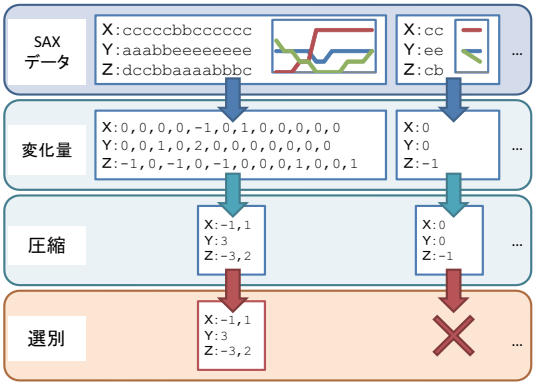

図 5 データの圧縮・選別の例

3.2 中間表現

テキスト生成では,時系列データと自然言語文をつなぐ中間表現を用いることでテキスト生成に使う言語資源を選択する.中間表現は表 1 のように定義する.

表 1 中間表現

action	中間表現	意味
up	“up{object}”	upward movement
down	“down{object}”	downward movement
left	“to_left{object}”	leftward movement
right	“to_right{object}”	rightward movement
right	“pass{object1,object2}”	cooperative movement

3.3 時系列データの動作判別

本研究では人の動作の判別を行うために対数線形モデルを用い,処理された時系列データと中間表現の対応を機械学習させる.3.1 で述べた時系列データ処理を施したデータ d と,人の動作を表す中間表現 y から構成した素性ベクトル ϕ を用いて,式 (1) の対数線形モデルを構成することで,データが与えられた下での各言語表現が選ばれる確率 $P(y|d)$ をモデル化した.ここで,$Z_{d,\mathbf{w}}$ は正規化係数である.

$$P(y|d) = \frac{1}{Z_{d,\mathbf{w}}} \exp(\mathbf{w} \cdot \phi(y, r)) \qquad (1)$$

3.4 バイグラムモデルによるテキスト生成

本研究では,バイグラムモデルを用いた単純なテキスト生成を行う.それぞれの動作に対しバイグラムモデルを構築するために被験者実験を行い,特定の動作に対して様々な自然言語表現を集めた.これにより,観測された時系列データに対して特定の中間表現が与えられたとき,言語資源としてバイグラムモデルを選択しテキスト生成を行う.しかし,例えば同じ動作でも,ある人は 10 語で表現し,またある人は 15 語で表現するなど,複数の表現方法がある.このことから,文の長さに依存しないテキスト生成が行えるよう,バイグラムモデルに null ラベルを導入する.null ラベルは,文の中の単語として扱われ,他の単語と同じようにユニグラムとバイグラムの構成要素となる.このように null ラベルを扱うために,構成されたバイグラムモデルに対して動的計画法を適用する前に以下に続く前処理をそれぞれの文に対して行う.まず,全ての文で単語数の最大値 max,最小値 min を得る.次に,max から min を引き,null に振る番号の最大値 $null\_max$ を求める.最後に,それぞれの文に対し,単語数が max に満たなければ,$null\_max$ から 1 ずつ引いた値を,足りない数だけ文末から文頭に向け挿入していく.null ラベル導入のイメージを,図 6 に示す.

文中の各 null ラベルに違う番号をつけることによって別の単語として見なし,それぞれがバイグラムモデルの 1 要素として扱う.また,本研究ではバイグラムモデルを構築する際に,使用する文の取捨選択を行わないことで,多くの語と関連づけることができるため,より複雑なテキスト生成を行うことができる.人の動作「*pick up ball*」を説明したバイグラムモデルのイメージを図 7 に示す.

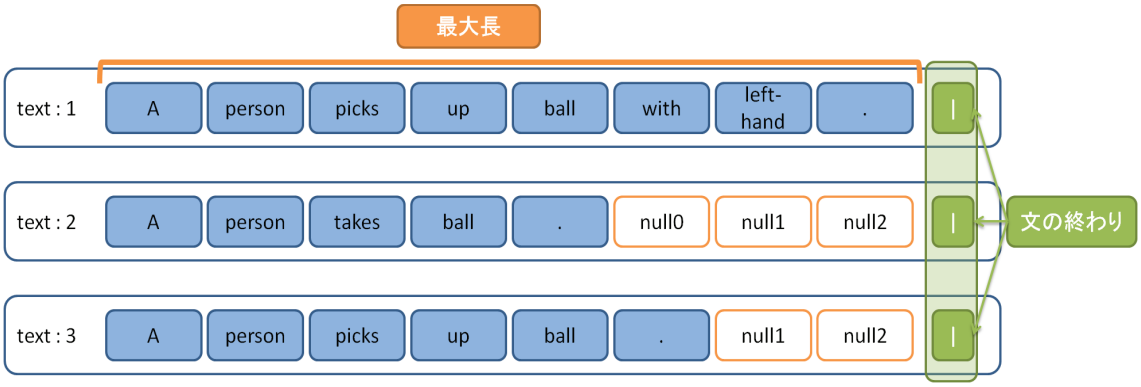

図 6 null ラベル導入のイメージ

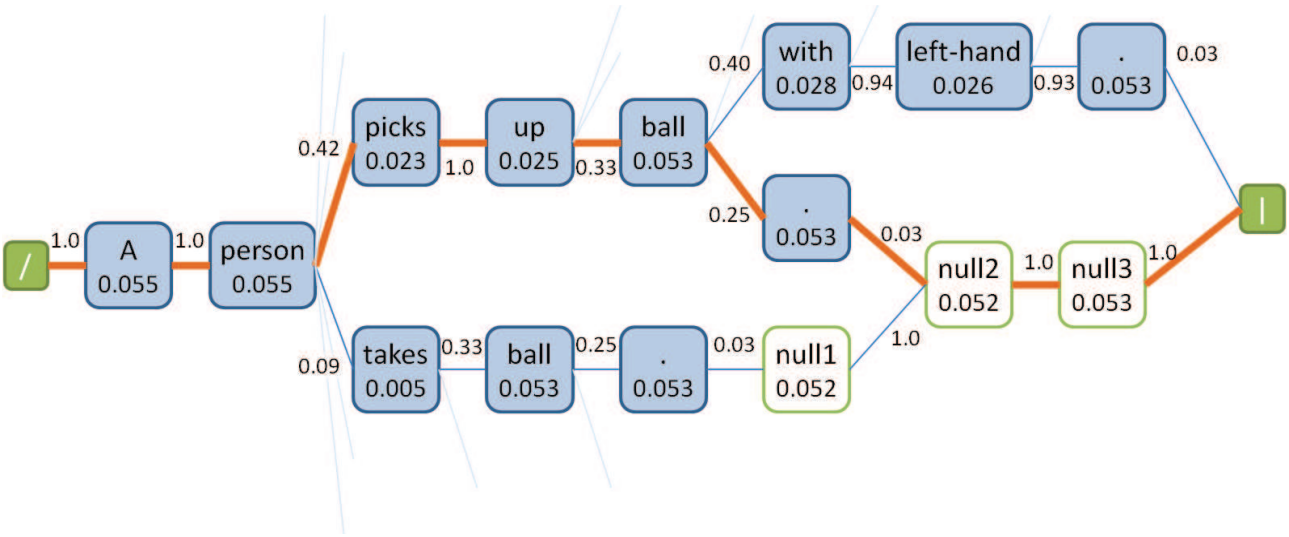

図 7 null ラベルを用いたバイグラムのイメージ

人の動作を説明するのによく用いられる文を生成するためには，このバイグラムモデルに動的計画法を適用することで尤度が最も高くなる単語の組み合わせからなる文を選ぶ．

4 実験

ここでは,「ボールを持ち上げて箱に入れる」という簡単な動作 (図 8) を言葉で表現することを目的とする．

図 8 言語化の対象となる動作

4.1 実験仕様

動作のどの部分を自然言語文で説明するのかを自動で決定することが難しいため、ここでは、言語化の対象となる動作を「*pick*」「*pass*」「*put*」の 3 つの基本動作から成ると定義するここでは，それぞれの動作に対し，自然言語での説明文を生成することとする．被験者実験として，対象となる人の動作の Kinect ビデオを観賞し，その動作について自然言語で説明してもらうという実験を 12 人に対し行った．収集した日本語の説明文を英訳し，各文に始端記号「/」と終端記号「|」を付加したものを言語資源としてバイグラムモデルを構築した．言語資源となった英文の全文数，語数，語の種類数を表 2,「*put*」におけるユニグラムとバイグラムを表 3 に示す．

動作判別には 3.3 で示した対数線形モデルを適用し，テキスト生成に使われる中間表現の判別に用いた．この識別器は，対象となる動作を捉えた 20 のデータを 15 の訓練データと 5 の評価データに分割し 5 クロスバリエーションした結果，精度の平均は 84%となった．

表 4 各動作に対する生成文の上位 3 文

動作	生成文	尤度
1	• A, person, picks, up, pink, ball, . , null_8, null_9, null_10, null_11, null_12, null_13, null_14, null_15, |	5.68e-24
	• A, person, picks, up, ball, with, left-hand, . , null_8, null_9, null_10, null_11, null_12, null_13, null_14, null_15	2.52e-24
	• A, person, picks, up, pink, ball, with, left-hand, . , null_8, null_9, null_10, null_11, null_12, null_13, null_14	2.10e-24
2	• A, person, passes, ball, to, right-hand, . , null_7, null_8, null_9, null_10, null_11	6.29e-16
	• A, person, passes, red, ball, to, right-hand, . , null_7, null_8, null_9,null_11, null_10	3.08e-18
	• A, person, passes, ball, from, left, to, right-hand, . , null_7, null_8, null_9	2.05e-18
3	• A, person, puts, ball, in, box, . , null_8, null_9, null_10, |	4.90e-15
	• A, person, puts, ball, in, box, . , null_7, null_8, null_9, null_10	1.22e-15
	• A, person, puts, ball, to, another, box, . , null_7, null_8, null_9	2.16e-16

表 2 収集された文の特徴

動き	文数	語数	語の種類数
pick	33	214	47
pass	18	148	28
put	36	290	28

4.2 実験結果

構築した識別機を用いて入力された動画から判別された中間表現は，1 つ目の動作は “up{object}” ，2 つ目の動作は “pass{object1,object2}” ，3 つ目の動作は “down{object}”’ と認識された．次に，選ばれた中間表現に対してあらかじめ構築されたバイグラムモデルに動的計画法を適用することで，動作を説明するもっともらしい文を生成する．

結果として，それぞれの動作に対して尤度の高かった上位 3 文を表 4 に示す．

4.3 考察

実験結果から，人の動作を正確に表現する文が生成出来ていることが確認できた．また，表 4 の生成文をみると，いくつかの文で終端文字 「|」 が出てきていないことが分かる．これは，バイグラムモデルが集めた文に現れる語のバイグラムの組み合わせによって構成されているためである．これにより，バイグラムモデルへ null ラベルを加えた文が，集められたどの文よりも長く生成される可能性がある．また一方で，文が長くなればなるほど，その文の尤度が低くなっていく．したがって，集められた文より長い文は生成されないという仮定の下で，集めた文の最大の単語数を生成文の単語数とした．

表 3 「*put*」におけるユニグラム，バイグラム

語	ユニグラム	バイグラム	
		次の語	遷移確率
.	0.0769	• null8	0.4444
	0.0769	• null9	0.1111
	0.0769	• null10	0.0555
	0.0769	• |	0.0555
	0.0769	• null7	0.3333
/	0.0769	• A	1.0
A	0.0769	• person	1.0
another	0.0277	• box	0.9230
	0.0277	• place	0.0769
ball	0.0769	• to	0.2222
	0.0769	• into	0.25
	0.0769	• with	0.0833
	0.0769	• in	0.4444
box	0.0726	• to	0.0588
	0.0726	• with	0.1470
	0.0726	• .	0.7941
drops	0.0021	• ball	1.0
hand	0.0021	• .	1.0
in	0.0341	• box	0.5
	0.0341	• right	0.25
	0.0341	• another	0.25
into	0.0192	• box	0.5555
	0.0192	• right	0.3333
	0.0192	• another	0.1111
left-hand	0.0021	• .	1.0
moves	0.0213	• pink	0.2
	0.0213	• ball	0.8
null10	0.0726	• |	1.0
null7	0.0256	• null8	1.0
null8	0.0598	• null9	1.0
null9	0.0683	• null10	1.0
other	0.0021	• hand	1.0
person	0.0769	• drops	0.0277
	0.0769	• put	0.0277
	0.0769	• returns	0.0277
	0.0769	• moves	0.2777
	0.0769	• puts	0.6388
pink	0.0042	• ball	1.0
place	0.0021	• .	1.0
put	0.0021	• ball	1.0
puts	0.0491	• ball	1.0
returns	0.0021	• ball	1.0
right	0.0170	• box	1.0
right-hand	0.0128	• .	1.0
to	0.0213	• box	0.1
	0.0213	• right	0.1
	0.0213	• another	0.8
with	0.0170	• right-hand	0.75
	0.0170	• other	0.125
	0.0170	• left-hand	0.125

5 まとめと今後の課題

本研究では，動画像中の人の動作を表現する確率的言語生成の枠組みを提案した．Kinect ビデオで抽出された人の動作は，時系列データとして取得され，いくつかの次元圧縮手法を適用することで機械学習に適した形に変換される．また観測された人の動きを表現するために，被験者実験によって集められた自然言語文に基づきバイグラムモデルを構築し，動的計画法を適用することで，もっともらしい語の組み合わせを取得する．さらに，バイグラムモデルに番号を付けた null ラベルを導入することにより，文生成に単語数の制限をつけずに自然言語文生成を行うことができた．また，提案手法はテンプレートによるテキスト生成ではなく，確率的なモデルによる生成であることから，例えばさらに文を収集すればそれに合わせて出力文も変化していくなど，資源となる文書によって様々な自然言語表現を得ることができる．

一方で，現段階では構文制約や物体認識を取り入れてはいない．そのため今後の課題として，こうした知識を導入するとともに，より正確にイベントを説明するようなテキスト生成が行えるよう発展させていきたいと考える．また，中間表現とバイグラムモデルとの対応付けをより柔軟したり，一連の動作から自然言語文によって説明される動作を区切る問題にも取り組んでいきたい．

参考文献

[1] Ding, Duo and Metze, Florian and Rawat, Shourabh and Schulam, Peter F. and Burger, Susanne, 2012. Generating natural language summaries for multimedia, Proceedings of the Seventh International Natural Language Generation Conference, INLG '12, Utica, Illinois, pp.128–130

[2] Ding, Duo and Metze, Florian and Rawat, Shourabh and Schulam, Peter Franz and Burger, Susanne and Younessian, Ehsan and Bao, Lei and Christel, Michael G. and Hauptmann, Alexander, 2012. Beyond audio and video retrieval: towards multimedia summarization Proceedings of the 2nd ACM International Conference on Multimedia Retrieval, No.2, pp.1–8

[3] Tan, Chun Chet and Jiang, Yu-Gang and Ngo, Chong-Wah, 2011. Towards textually describing complex video contents with audio-visual concept classifiers, Proceedings of the 19th ACM international conference on Multimedia, pp.655–658

[4] Ichiro Kobayashi, Mami Noumi, and Atsuko Hiyama, 2010. A Study on Verbalization of Human Behaviors in a Room FUZZ-IEEE 2010 Barcelona, Spain, 18-23 July

[5] A. Barbu, A. Bridge, Z. Burchill, D. Coroian, S. Dickinson, S. Fidler, A. Michaux, S. Mussman, S. Narayanaswamy, D. Salvi, L. Schmidt, J. Shangguan, J. Siskind, J. Waggoner, S. Wang, J. Wei, Y. Yin, and Z. Zhang. 2012. *Video In Sentences Out*, Conference on Uncertainty in Artificial Intelligence (UAI),

[6] Mirella Lapata, 2003. Probabilistic Text Structuring: Experiments with Sentence Ordering, In Proc. of the Annual meeting of the Association for Computational Linguistics pp.545–552

[7] Anja Belz, 2007. Probabilistic Generation of Weather Forecast Texts, Proceedings of NAACL HLT 2007, pp.164-171

[8] Anja Belz and Eric Kow, 2009. System building cost vs. output quality in data-to-text generation, In Proceedings of the 12th European Workshop on Natural Language Generation (ENLG-09, pp.16-24, Athens, Greece

[9] Wei Lu and Hwee Tou Ng and Wee Sun Lee, 2009. Natural language generation with tree conditional random fields, EMNLP, pp.400–409

[10] Percy Liang, Michael I. Jordan, Dan Klein 2009. Learning Semantic Correspondences with Less Supervision, ACL-IJCNLP

[11] Angeli, Gabor and Liang, Percy and Klein, Dan, 2010. A simple domain-independent probabilistic approach to generation, Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pp. 502–512,Cambridge, Massachusetts

[12] Konstas, Ioannis and Lapata, Mirella, 2012. Unsupervised concept-to-text generation with hypergraphs, Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Montreal, Canada, pp.752–761

[13] Konstas, Ioannis and Lapata, Mirella, 2012. Concept-to-text generation via discriminative reranking, booktitle = Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1, pp. 369–378, Jeju Island, Korea

[14] Lin, J., Keogh, E., Lonardi, S. and Chiu, B. 2003. A Symbolic Representation of Time Series, with Implications for Streaming Algorithms DMKD' 03

[15] Srinivas Bangalore and Owen Rambow 2000. Exploiting a Probabilistic Hierarchical Model for Generation, Proceedings of the 18th conference on Computational Linguistics (Coling 2000), Volume 1, pp.42-48,