



EMNLP 2012 参加報告

<http://www.chokkan.org/publications/emnlp2012.pdf>

東北大学大学院情報科学研究科

東北大学工学部情報知能システム総合学科

東北大学電気通信情報機構

JST 戦略的創造研究推進事業「さきがけ」・「情報環境と人」

岡崎 直観 (okazaki at ecei tohoku ac jp)

<http://www.chokkan.org/>
@chokkanorg



EMNLP 2012

- スケジュール

- 2012年3月28日: 論文投稿締切り (ACL 2012の採択通知後)
- 2012年5月1日～4日: 著者回答期間
- 2012年5月18日: 採択通知
- 2012年6月3日: 最終原稿締切り
- 2012年7月12日～14日: 会議開催 (韓国の済州島, ACL2012から連続)

- オーガナイザ

- General Chair: Jun'ichi Tsujii
- Program Chairs: James Henderson
Marius Pasca
- Publication Chair: Naoaki Okazaki
- Area Chairs: Hiroya Takamura 他
- Program Committee (reviewers): 591名



Publication Chairのお仕事

- 5月下旬: ACL 2012のPublication Chairsと段取りを相談
- 5月下旬: START上のACLPubの使い方を習得
- 6月3日～20日: 怒濤の確認・編集作業
 - 提出されたcopyrightフォームがきちんと書かれているか確認
 - A4サイズになっていた論文への対応, 論文の余白の調整
 - STARTに登録されているメタデータのチェック
 - 論文PDFのタイトルとSTART上のタイトルが違う! $\Sigma(^{\circ}\Delta^{\circ})$ ガーン
 - 論文中の著者名とSTART上の著者名が違う! $\Sigma(^{\circ}\Delta^{\circ}\text{III})$ ガーン
 - 前書きのプランニング, 原稿の取り立て, 編集
 - 会議のプログラムのチェック作業 (6月13日のメールやりとりだけで92回)
 - 最終版のPDFの生成

Publication Chairをやってみて思ったこと

- すごい研究者は何にでも全力！
 - ACLPub(プロシーディング生成システム)はDavid Yarowskyのスクリプトを基に, Jason EisnerとPhilipp Koehnが2005年に制作
- 貴重な資料がいっぱい(古くなっているものもありますが)
 - Jason Eisner. 2007. [How to Serve as Program Chair of a Conference.](#)
 - Joakim Nivre and Noah Smith. 2008. [How to Serve as Publications Chair for an ACL Conference.](#)
 - [Conference Handbook \(会議・ワークショップ開催に関する資料\)](#)
- とにかくメールを送ることは重要
 - 発生するトラブルのほとんどはメールを出さない, もしくは(意図せず)受信者から外れてしまっていることによるコミュニケーション不足
 - みんな忙しいので, 気づいたことはどんどんメールする

会議のハイライト

- Invited Talks

- Eric Xing (Carnegie Mellon University). [On Learning Sparse Structured Input-Output Models](#).
- Patrick Pantel (Microsoft Research). The Appification of the Web and the Renaissance of Conversational User Interfaces

- Best paper awards

- Annie Louis and Ani Nenkova (UPenn) A coherence model based on syntactic patterns.
- David Hall and Dan Klein (University of California, Berkeley). Training Factored PCFGs with Expectation Propagation.
- Ndapandula Nakashole, Gerhard Weikum and Fabian Suchanek (Max Planck Institute for Informatics). PATTY: A Taxonomy of Relational Patterns with Semantic Types

(Best paper)

A coherence model based on syntactic patterns

Annie Louis and Ani Nenkova (UPenn)
EMNLP 2012, pages 1157-1168

背景

- テキストの一貫性のモデル化の研究において「意図構造」が注目されてこなかった

仮説

- 統語構造から取り出される素性（生成規則など）が意図構造や一貫性を映し出すのでは？

貢献

- 隣接する文間の統語素性のペアの中には、確かに意図構造を反映するものが存在しそうだ
- 統語素性からテキストの一貫性をモデル化する手法（局所確率モデル, HMM）を提案
- 統語素性がテキストの一貫性, 文の意図構造, 文章のタイプなどを予測できることを実証した

テキストの一貫性 (Grosz and Sidner, 86)

グロス

サイドナー

注意構造

言語構造

意図構造

- 注意状態 (attentional structure)
 - 談話の進行にともなう注意の焦点
 - 「どんな内容が議論されているか」
- 言語構造 (linguistic structure)
 - 談話単位 (discourse segment) の列を組み上げたもの
- 意図構造 (intentional structure)
 - 談話の背後にある意図の構造
 - 「概念・事態を説明したい」「批判したい」など

既存研究: テキストの一貫性

注意構造

言語構造

意図構造

- Entity Grid (固有表現の繰り返しにおける統語的役割の推移)
 - (Barzilay and Lapata, 08; Elsner and Charniak, 08)

	Department	Trial	Microsoft	Evidence	Competitors	Markets	Products	Brands	Case	Netscape	Software	Tactics	Government	Suit	Earnings	
1	s	O	s	X	O	-	-	-	-	-	-	-	-	-	-	1
2	-	-	O	-	-	X	s	O	-	-	-	-	-	-	-	2
3	-	-	s	O	-	-	-	-	s	O	O	-	-	-	-	3
4	-	-	s	-	-	-	-	-	-	-	-	s	-	-	-	4
5	-	-	-	-	-	-	-	-	-	-	-	-	s	O	-	5
6	-	X	s	-	-	-	-	-	-	-	-	-	-	-	O	6

S: subject; O:object; X:other position
Barzilay and Lapata (2008) より

一貫性のあるテキストでは密な列 (例: Microsoft) と疎な列 (例: Markets) が見られる

密な列 (例: Microsoft) ではSやOが出やすい

隣り合う文同士からgridの値の遷移確率 (例えば "S-" が 0.08 など) を素性とするれば, テキストの一貫性を評価 (機械学習によるランキング) できる

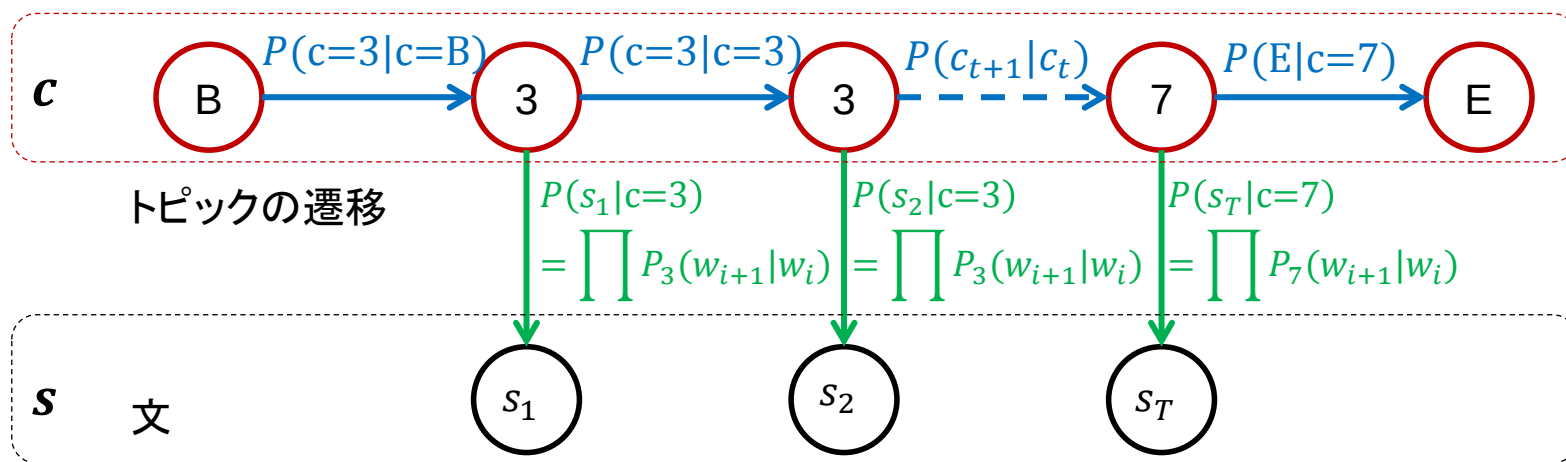
既存研究: テキストの一貫性

注意構造

言語構造

意図構造

- HMMによるトピックの遷移と文(単語列)の生成のモデル化
 - (Barzilay and Lee, 04; Fung and Ngai, 06)



- コーパス中の文をクラスタリングしてトピックを表すクラスター $\{c_1, \dots, c_m\}$ を作る
- 各クラスタをHMMにおける「状態」と考える
 - 状態遷移確率 $P(c_{t+1}|c_t)$: c_t に属する文に続き c_{t+1} に属する文が出現する条件付き確率
 - 文生成確率 $P(s_t|c_t)$: クラスター c_t に属する文集合におけるバイグラム言語モデル
- テキストの生成確率 $P(S) = \prod P(s_{t+1}|s_t)$ を求め、テキストの一貫性を評価する

既存研究: テキストの一貫性

注意構造

言語構造

意図構造

- 談話関係は読みやすさと関連が深い (Pitler and Nenkova, 08)
- 分析対象: PDTBの記事に読みやすさスコアを人手で付けたもの
- 2つのタスクに貢献する素性を分析
 - 線形回帰モデルによる読みやすさのスコアの予測
 - (2つの記事が与えられた時) 線形SVMによる読みやすさの優劣の予測
- 以下の素性セットに対して, 素性の貢献を調べた
 - テキストの長さ
 - WSJコーパスから作った言語モデルに対する尤度
 - 統語素性(構文木の高さ, 平均NP数/文, 平均VP数/文, 平均従属節数/文)
 - 語彙的結束性素性(隣り合う文での名詞の重なり数, コサイン類似度など)
 - Entity Grid(固有表現の出現パターン)による一貫性
 - PDTBから作った談話関係の「言語モデル」に対する尤度 ← 両方のタスクで安定して貢献

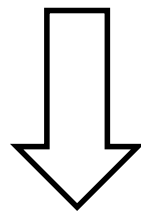
テキストの一貫性

注意構造

言語構造

意図構造

- 意図構造
 - 談話の背後にある話し手の意図の構造
 - これまでほとんど研究されてこなかった



本研究

統語的なパターンが意図構造の
代用品として有用

一体どういうアイデアか？

導水管

An **aqueduct** is a [water](#) supply or navigable [channel](#) constructed to convey water. In modern engineering, the term is used for any system of pipes, ditches, canals, tunnels, and other structures used for this purpose.^[1]

サイトカイン受容体

Cytokine receptors are [receptors](#) that bind [cytokines](#).^[1] In recent years, the cytokine receptors have come to demand the attention of more investigators than cytokines themselves, partly because of ...

- 先頭の文は定義文の典型

- NPの後にcopular verb (is/are) が続き、導入したい実体・概念を説明
- 補語は以下の2つの部分から構成される
 - 導入した実体・概念の親の概念を説明するNP
 - 説明したい実体・概念に固有な属性を説明する従属節

- 2番目の文で実体・概念をさらに説明

- Topicalized phraseから始まることにより、2番目の文の焦点を定める

生成規則により示される談話意図の例

VP → VB VP

"The refund pool may not [be held hostage through another round of appeals] **VP**," Judge Curry said.

「VP → VB VP」のVBは
モーダルと結びつきやすい

推量 → 支持

固有名詞が主語になる

NP-SBJ → NNP NNP

[Commonwealth Edison] **NP-SBJ** said it is already appealing the underlying commission order and is considering appealing Judge Curry's order.

NP-LOC → NNP

"It has to be considered as an additional risk for the investor," said Gary P. Smaby of Smaby Group Inc., [Minneapolis] **NP-LOC**.

「NP-LOC → NNP」は場所
から人のタイトルを表す

Xを紹介 → Xが述べる

トピック構造

S-TOC-1 → NP-SBJ VP

["Cray Computer will be a concept stock,"] **S-TPC-1** he said.

予備実験: 統語素性による一貫性の例

- Penn TreebankのSection 0(99文書, 1727文)において
 - 出現回数が25回未満の生成規則は削除 → 197個の生成規則に
 - 隣接文対において χ^2 検定で統計的に有意に出現する生成規則のペアを列挙
- 顕著に出現するペアの特徴
 - 先行研究で指摘されている同一の生成規則の繰り返しも見つかるが, 全体の5%くらいで少ない
 - 数量の説明など, 談話の意図が見えやすいペアも多く見つかる

p_1	p_2	$c(p_1p_2)$
	— Repetition —	
VP → VBD SBAR	VP → VBD SBAR	83
QP → \$ CD CD	QP → \$ CD CD	18
NP → \$ CD -NONE-	NP → \$ CD -NONE-	16
NP → QP -NONE-	NP → QP -NONE-	15
NP-ADV → DT NN	NP-ADV → DT NN	10
NP → NP NP-ADV	NP → NP NP-ADV	7
	— Quantities/Amounts —	
NP → QP -NONE-	QP → \$ CD CD	16
QP → \$ CD CD	NP → QP -NONE-	15
NP → NP NP-ADV	NP → QP -NONE-	11
NP-ADV → DT NN	NP → QP -NONE-	11
NP → NP NP-ADV	NP-ADV → DT NN	9
NP → \$ CD -NONE-	NP-ADV → DT NN	8
NP-ADV → DT NN	NP → \$ CD -NONE-	8
NP-ADV → DT NN	NP → NP NP-ADV	8
NP → NP NP-ADV	QP → CD CD	6
	— Other —	
S → NP-SBJ VP	NP-SBJ → PRP	290
VP → VBD SBAR	PP-TMP → IN NP	79
S → NP-SBJ-1 VP	VP → VBD SBAR	43
VP → VBD NP	VP → VBD VP	31
VP → VB VP	NP-SBJ → NNP NNP	27
NP-SBJ-1 → NNP NNP	VP → VBD NP	13
VP → VBZ NP	S → PP-TMP, NP-SBJ VP.	8
NP-SBJ → JJ NNS	VP → VBP NP	8
NP-PRD → NP PP	NP-PRD → NP SBAR	7
NP-LOC → NNP	S-TPC-1 → NP-SBJ VP	6

Table 3: Top patterns in productions from WSJ

手法の準備: 統語素性の2種類の表現形式

- 生成規則

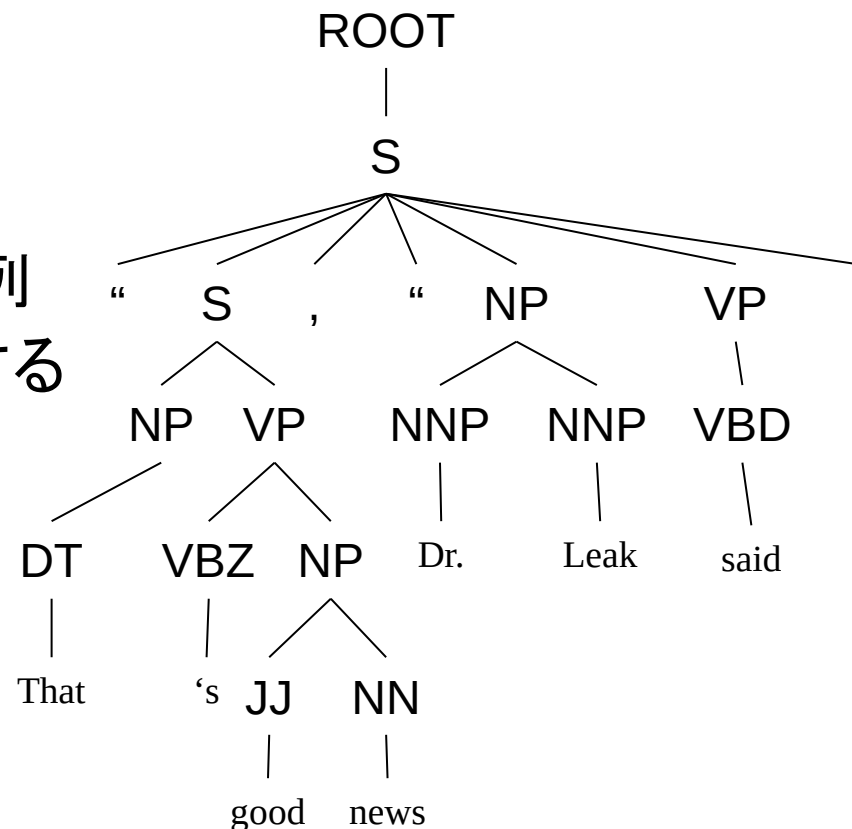
- 終端記号(単語)は除去

- d*-sequence

- 深さ*d*における非終端記号の系列
- 最も左下の非終端記号で修飾する

- 句の内部構造を表現できる
 - “the chairs”: NP_{DT}
 - “his chairs”: NP_{PRP\$}
 - “comfortable chairs”: NP_{JJ}

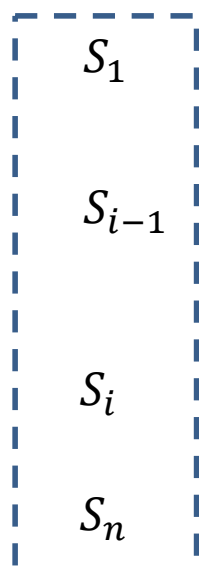
- 深さ2: “ S_{DT} , “ NP_{NNP} VP_{VBD} .
- 深さ3: “ NP_{DT} VP_{VBZ} , “ NNP NNP VBD .



手法1: 局所的な一貫性モデル

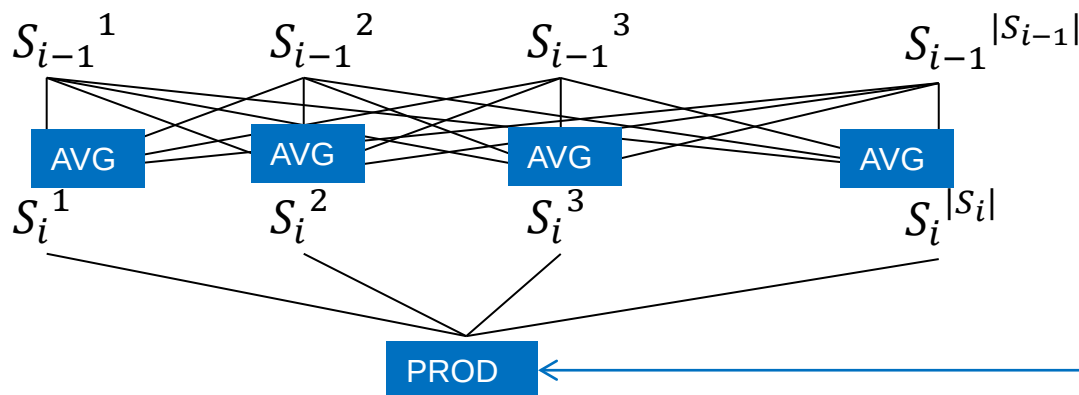
- テキスト T の一貫性を次式で測る

$$P(T) = \prod_{i=2}^n \left[\prod_{j=1}^{|S_i|} \frac{1}{|S_{i-1}|} \sum_{k=1}^{|S_{i-1}|} p(S_i^j | S_{i-1}^k) \right]$$



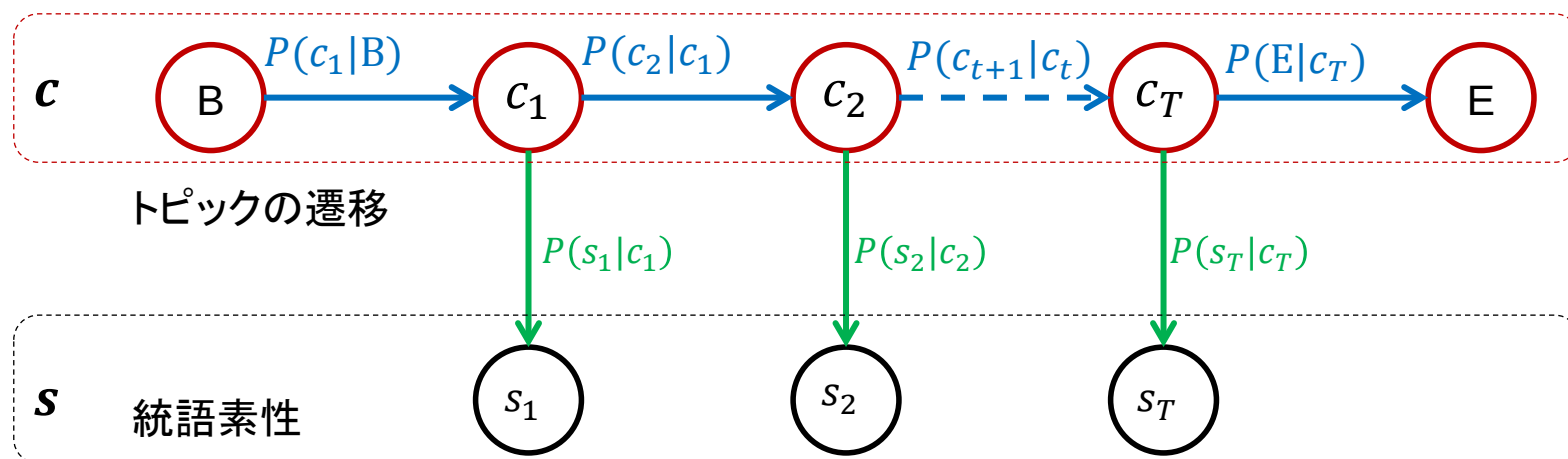
テキスト T : 文の系列

隣り合う文が含む統語素性の全ペアの条件付き確率
(スムージング有)から $S_{i-1} \rightarrow S_i$ の遷移確率を計算



手法2: HMMによる一貫性モデル

- Barzilay and Lee (04) のアプローチを踏襲
 - 文を生成するモデル → **統語素性を生成するモデル** に変更
 - HMMの状態: 統語素性のコサイン類似度を用いて文をクラスタリング
 - 状態遷移確率 $P(c_{t+1}|c_t)$: c_t に属する文に続き c_{t+1} に属する文が出現する確率
 - 文生成確率 $P(s_t|c_t)$: クラスタ c_t に属する文集合から構築した「統語素性生成言語モデル」
 - 生成規則の場合: それぞれの生成規則を生成する確率(ユニグラムモデル)
 - d-sequenceの場合: d-sequence中のbi-gramを生成する確率



評価実験: 一貫性のあるテキストの選択

Model	Accidents		Earthquake	
	Parameter	Acc	Parameter	Acc
A. Local co-occurrence				
Prodns		72.8		55.0
<i>d</i> -seq	dep. MVP+2	71.8	dep. MVP+1	65.1
POS		61.3		42.6
B. HMM-syntax				
Prodns	clus. 37	74.6	clus. 5	93.8
<i>d</i> -seq	dep. MVP+8	82.2	dep. MVP+9	86.5
	clus. 8		clus. 45	
C. Other approaches				
Egrid	history 1	67.6	history 1	82.2
Content	clus. 48	71.4	clus. 23	84.5

ランダムに並び替えた文章と並び替えていない文章を与え、後者を選ぶタスク

統語素性からテキストの一貫性を予測できた

局所的な一貫性モデルよりもHMMによる一貫性モデルの方が性能が良かった

既存手法よりも提案手法の方が高精度

Table 5: Accuracies on accident and earthquake corpora

評価実験: 複数の手がかりの組み合わせ

Model	※太字は組み合わせ で改善が見られたもの	Accid.	Earthq.
Content + Egrid	→	76.8	90.7
Content + HMM-prodn		74.2	95.3
Content + HMM- <i>d</i> -seq		82.1	90.3
Egrid + HMM-prodn		79.6	93.9
Egrid + HMM- <i>d</i> -seq		84.2	91.1
Egrid + Content + HMM-prodn		79.5	95.0
Egrid + Content + HMM- <i>d</i> -seq		84.1	92.3
Egrid + Content + HMM-prodn + HMM- <i>d</i> -seq		83.6	95.7

※2値分類問題としてロジスティック回帰で複数の素性を統合

Table 6: Accuracies for combined approaches

意図構造と統語素性クラスタの関連づけ

- ART Corpus(225件の化学分野の論文概要)でアノテートされた意図構造に対応するクラスタを作ることが出来るか？
 - 意図構造: 結果, 結論, 目的, 手法, 背景, 所見, 実験, 動機, モデル, 仮説
 - クラスタと意図構造の相関が強いペアを χ^2 検定で結びつける

Abstracts (HMM-prod 9 clusters)			
Positive associations	matches	prec.	recall
Clus5 - Model	7	17.1	43.8
Clus7 - Objective	27	27.6	32.9
Clus7 - Goal	16	16.3	55.2
Clus0 - Conclusion	15	50.0	12.1
Clus6 - Conclusion	27	51.9	21.8
Not associated: Clus7 - Conclusion, Clus8 - Conclusion			

Introductions (HMM-d-seq 6 clusters)			
Positive associations	matches	prec.	recall
Clus2-Background	161	64.9	14.2
Clus3-Objective	37	7.9	38.5
Clus4-Goal	29	9.8	32.6
Clus4-Hypothesis	12	4.1	52.2
Clus5-Motivation	61	12.9	37.4
Not associated: Clus1 - Motivation, Clus2 - Goal, Clus4 - Background, Clus 5 - Model			

- 意図構造と相関が強いペアを見つけることができる
 - (ちょっと厳しいような…)

論文コーパスにおけるその他の実験

- ランダムに並び替えた文章と、元々の並び順の文章が与えられ、元々の並び順の文章を選ぶ実験

Data	Section	Test pairs	Local-prod	Local-d-seq	HMM-prod	HMM-d-seq	Oracle zones
ART Corpus	Abstract	1633	57.0	52.9	64.1	55.0	80.8
	Intro	1640	44.5	54.6	58.1	64.6	94.0
ACL Conference	Abstract	8815	44.0	47.2	58.2	63.7	
	Intro	9966	54.5	53.0	64.4	74.0	
	Rel. wk.	10,000	54.6	54.4	57.3	67.3	

Table 8: Accuracy in differentiating permutation from original sections on ACL and ART test sets.

- ACLの本会議の論文とワークショップの論文を分類できるか？

- 本会議: 完成された研究
- ワークショップ: 途中の研究
- 二値分類問題として定式化

Conf.	Abstract	Intro	Rel wk
≥ 0.5	59.3 (100.0)	50.3 (100.0)	55.4 (100.0)
≥ 0.6	63.8 (67.2)	50.8 (71.1)	58.6 (75.9)
≥ 0.7	67.2 (32.0)	54.4 (38.6)	63.3 (52.8)
≥ 0.8	74.0 (10.0)	51.6 (22.0)	63.0 (25.7)
≥ 0.9	91.7 (2.0)	30.6 (5.0)	68.1 (7.2)

- アブストラクトを使うと判別が可能
- 論文の意図構造のモデル化が可能

Table 9: Accuracy (% examples) above each confidence level for the conference versus workshop task.

感想など

- 統語素性だけで一貫性がモデル化出来るというのは興味深い
 - 語彙素性と統語素性を組み合わせられるようになると楽しそう
- 談話解析器 (Lin+, 09; Feng and Hearst, 12) に構文木の生成規則を素性として入れる意義がクリアになった
- 隣接文を考慮した構文解析 (Cheung and Penn, 10) との関連
- 一貫性は確かにモデル化出来てそうであるが、モデル化出来ているものが「意図構造」なのか、検証が不十分

その他

その他面白かった研究

- D Hall and D Klein. Training Factored PCFGs with Expectation Propagation
 - PCFGの文法規則をヘッド単語, 親のシンボル, 隠れ変数など, 複数のアノテーションを混ぜて使えるようにした
- V Chahuneau, K Gimpel, B R Routledge, L Scherlis and N A Smith. Word Salad: Relating Food Prices and Descriptions
 - メニューの記述から食べ物の値段を予測する
 - 手法の目新しさはないが, とにかくプレゼンが楽しい
- A Goyal, H Daume III and R Guerra. Fast Large-Scale Approximate Graph Construction for NLP
 - Online PMIで単語共起行列を近似的に求める
 - Count Min sketchで類似頻度計測と並列化を同時に実現
 - [解説資料](#)あり

国際会議スケジュール (※非公開・誤りの情報を含む)

会議名	開催場所	投稿〆切	開催日
PAKDD 2013	Gold Coast, Australia	2012/10/1	2013/4/14-17
WWW 2013	Rio de Janeiro, Brazil	2012/11/26	2013/5/13-17
NAACL 2013	Atlanta, USA	2012/12/7	2013/6/9-14
ICML 2013	Atlanta, USA	2012/10/1	2013/6/16-21
ICDM 2013	New York, USA	2012/12/18	2013/6/16-21
SIGIR 2013	Dublin, Ireland	2013/1/28	2013/7/28-8/1
IJCAI 2013	Beijing, China	?	2013/8/3-9
ACL 2013	Sofia, Bulgaria	(NAACL採択通知後)	2013/8/4-9
CoNLL 2013	Sofia, Bulgaria	?	2013/8/8-9(?)
KDD 2013	Chicago, USA	?	2013/8/11-14
EMNLP 2013	(North America)	(ACL採択通知後)	2013年秋
IJCNLP 2013	Nagoya, Japan	?	2013/10/15-17
EACL 2014	Göteborg, Sweden	?	2014/4/26-30