Learning Compression & Linguistic Quality

Ling 573
Systems and Applications
May 14, 2015

Roadmap

- Sentence Compression:
  - Learning compression: Tree-based approach
  - Results & Discussion

- Linguistic Quality:
  - Corpus study and analysis
  - Automatic evaluation
  - Improvements for MDS

Tree-based Compression

- Given a phrase-structure parse tree,
  - Determine if each node is: removed, retained, or partial

Tree-based Compression

- Given a phrase-structure parse tree,
  - Determine if each node is: removed, retained, or partial
- Issues:

Tree-based Compression

- Given a phrase-structure parse tree,
  - Determine if each node is: removed, retained, or partial
- Issues:
  - # possible compressions exponential

Tree-based Compression

- Given a phrase-structure parse tree,
  - Determine if each node is: removed, retained, or partial
- Issues:
  - # possible compressions exponential
  - Need some local way of scoring a node

Tree-based Compression

- Given a phrase-structure parse tree,
  - Determine if each node is: removed, retained, or partial
- Issues:
  - # possible compressions exponential
  - Need some local way of scoring a node
  - Need some way of ensuring consistency
    - I.e. can’t have retain over remove
  - Need to ensure grammaticality

Tree-based Compression

- Given a phrase-structure parse tree,
  - Determine if each node is: removed, retained, or partial
- Issues & Solutions:
  - # possible compressions exponential
    - Order parse tree nodes (here post-order)
    - Do beam search over candidate labelings
  - Need some local way of scoring a node

Tree-based Compression

- Given a phrase-structure parse tree,
  - Determine if each node is: removed, retained, or partial
- Issues & Solutions:
  - # possible compressions exponential
    - Order parse tree nodes (here post-order)
    - Do beam search over candidate labelings
  - Need some local way of scoring a node
    - Use MaxEnt to compute probability of label
  - Need some way of ensuring consistency

Tree-based Compression

- Given a phrase-structure parse tree,
  - Determine if each node is: removed, retained, or partial
- Issues & Solutions:
  - # possible compressions exponential
    - Order parse tree nodes (here post-order)
    - Do beam search over candidate labelings
  - Need some local way of scoring a node
    - Use MaxEnt to compute probability of label
  - Need some way of ensuring consistency
    - Restrict candidate labels based on context
  - Need to ensure grammaticality

Tree-based Compression

- Given a phrase-structure parse tree,
  - Determine if each node is: removed, retained, or partial
- Issues & Solutions:
  - # possible compressions exponential
    - Order parse tree nodes (here post-order)
    - Do beam search over candidate labelings
  - Need some local way of scoring a node
    - Use MaxEnt to compute probability of label
  - Need some way of ensuring consistency
    - Restrict candidate labels based on context
  - Need to ensure grammaticality
    - Rerank resulting sentences using n-gram LM

Features

- Basic features:
  - Analogous to those for sequence labeling

Features

- Basic features:
  - Analogous to those for sequence labeling
- Enhancements:
  - Context features: decisions about child, sibling nodes

Features

- Basic features:
  - Analogous to those for sequence labeling
- Enhancements:
  - Context features: decisions about child, sibling nodes
  - Head-driven search:
    - Reorder so head nodes at each level checked first
      - Why?

Features

- Basic features:
  - Analogous to those for sequence labeling
- Enhancements:
  - Context features: decisions about child, sibling nodes
  - Head-driven search:
    - Reorder so head nodes at each level checked first
      - Why? If head is dropped, shouldn't keep rest
      - Revise context features

Summarization Features

  - (aka MULTI in paper)
- Calculated based on current decoded word sequence W
- Linear combination of:

Summarization Features

  - (aka MULTI in paper)
- Calculated based on current decoded word sequence W
- Linear combination of:
  - Score under MaxEnt
  - Query relevance:
    - Proportion of overlapping words with query
  - Importance: Average sumbasic score over W
  - Language model probability
  - Redundancy: 1 --- proportion of words overlapping summ

Summarization Results

	DUC 2006			DUC 2007		
System	**C Rate**	**R-2**	**R-SU4**	**C Rate**	**R-2**	**R-SU4**
Best DUC system	–	9.56	15.53	–	12.62	17.90
Davis et al. (2012)	–	10.2	15.2	–	12.8	17.5
SVR	100%	7.78	13.02	100%	9.53	14.69
LambdaMART	100%	9.84	14.63	100%	12.34	15.62
Rule-based	78.99%	10.62 ∗†	15.73 †	78.11%	13.18†	18.15†
Sequence	76.34%	10.49 †	15.60 †	77.20%	13.25†	18.23†
Tree (BASIC + $Score_{Basic}$)	70.48%	10.49 †	15.86 †	69.27%	13.00†	18.29†
Tree (CONTEXT + $Score_{Basic}$)	65.21%	10.55 ∗†	16.10 †	63.44%	12.75	18.07†
Tree (HEAD + $Score_{Basic}$)	66.70%	10.66 ∗†	16.18 †	65.05%	12.93	18.15†
Tree (HEAD + MULTI)	70.20%	**11.02** ∗†	**16.25** †	73.40%	**13.49**†	**18.46**†

Compression Results

System	C Rate	Uni-Prec	Uni-Rec	Uni-F1	Rel-F1
HedgeTrimmer	57.64%	0.72	0.65	0.64	0.50
McDonald (2006)	70.95%	0.77	0.78	*0.77*	0.55
Martins and Smith (2009)	71.35%	0.77	0.78	*0.77*	0.56
Rule-based	87.65%	0.74	0.91	0.80	0.63
Sequence	70.79%	0.77	0.80	*0.76*	0.58
Tree (BASIC)	69.65%	0.77	0.79	0.75	0.56
Tree (CONTEXT)	67.01%	**0.79**	0.78	*0.76*	0.57
Tree (HEAD)	68.06%	**0.79**	0.80	*0.77*	**0.59**

Discussion

- Best system incorporates:
  - Tree structure
  - Machine learning
  - Summarization features

Discussion

- Best system incorporates:
  - Tree structure
  - Machine learning
  - Summarization features
- Rule-based approach surprisingly competitive
  - Though less aggressive in terms of compression

Discussion

- Best system incorporates:
  - Tree structure
  - Machine learning
  - Summarization features
- Rule-based approach surprisingly competitive
  - Though less aggressive in terms of compression
- Learning based approaches enabled by sentence compression corpus

General Discussion

- Broad range of approaches:
  - Informed by similar linguistic constraints
  - Implemented in different ways:

General Discussion

- Broad range of approaches:
  - Informed by similar linguistic constraints
  - Implemented in different ways:
    - Heuristic vs Learned
    - Surface patterns vs parse trees vs SRL
  - Even with linguistic constraints
    - Often negatively impact linguistic quality

General Discussion

- Broad range of approaches:
  - Informed by similar linguistic constraints
  - Implemented in different ways:
    - Heuristic vs Learned
    - Surface patterns vs parse trees vs SRL
  - Even with linguistic constraints
    - Often negatively impact linguistic quality
    - Key issue: errors in linguistic analysis
      - POS taggers → Parsers → SRL, etc

General Discussion

- Compression has range of uses:
  - Removing irrelevant information for selection
  - Improving readability
  - Allowing inclusion of more information
- Slightly different strategies for each

Linguistic Quality

Evaluation

- Shared tasks:
  - Take content as primary evaluation measure
    - ROUGE, Pyramid, (manual) Responsiveness
  - Linguistic quality also part of formal evaluation
- TAC “Readability”:
  - Scored manually on 5-point Likert scale
  - Aims to capture readability, fluency
    - Independent of summary content

What is "Readability"?

- According to TAC,
- Assessors consider (and rate 1-5) each of:
  - Grammaticality:
    - No fragments, datelines, ill-formed sentences, etc
  - Non-redundancy:
    - No unnecessary repetition: includes content, sentences, or full NPs when pronoun is better
  - Referential clarity:
    - Both presence/salience of antecedents, relevance of items
  - Focus:
    - Only content related to summary
  - Coherence: "Well-structured"

What is “Readability”? II

- Definition subsumes many phenomena, errors
- What types of errors do these systems make?
- What errors, issues are reflected in the scores?
- LVQSumm (Friedrich et al, 2013)
  - Annotate linguistic “violations” in automatic summaries
    - TAC2011 data: 2000 “peer” summaries; GLOW system
  - Categorize and tabulate
  - Assess correlation with Readability scores

Example

Charles Carl Roberts IV may have planned to molest the girls at the Amish school, but police have no evidence that he actually did. Charles Carl Roberts IV entered the West Nickel Mines Amish School in Lancaster County and shot 10 girls, killing five. The suspect apparently called his wife from a cell phone shortly before the shooting began, saying ***he was “acting out in revenge for something that happened 20 years ago,*** *Miller said. The gunman, a local truck driver Charles Roberts, was apparently* ***acting in “revenge” for an incident that happened to him 20 years ago.***

Violation Categories

- Entity mentions:
  - Affect coreference and readability
  - 1st mention w/o explanation; subseq. Mention w/expl
  - Def NP w/o prev mention; indef NP w/ prev mention
  - Pron w/missing, misleading antecedent; Acronym
- Clausal level:
  - Arbitrary spans – up to sentence level
  - Incomplete sent, dateline info, other ungrammatical
  - No semantic relation, wrong discourse rel'n, redundancy

violation type	count	avg/doc	Pearson's *r* Readability	Pyramid	Respons.
entity level violations					
DNP-REF	958	0.50	**-0.122**	**-0.166**	**-0.133**
FM-EXPL	792	0.41	0.006	-0.050	**-0.066**
INP+REF	430	0.22	-0.052	**0.235**	**0.109**
PRN+MISSA	361	0.19	**-0.191**	**-0.140**	**-0.156**
SM+EXPL	162	0.08	0.020	**0.089**	0.045
PRN+MISLA	27	0.01	**-0.065**	**-0.073**	**-0.089**
ACR-EXPL	11	0.01	-0.038	-0.056	-0.006
sum(DNP-REF, PRN+MISSA)	1319	0.68	**-0.204**	**-0.208**	**-0.192**
sum(entity level violations)	2741	1.42	**-0.167**	-0.074	**-0.127**
clause level violations					
INCOMPLSN	1,044	0.54	**-0.210**	0.000	-0.029
OTHRUNGR	793	0.41	**-0.180**	0.007	-0.016
INCLDATE	412	0.21	**-0.090**	0.039	0.051
REDUNDINF	504	0.26	**-0.160**	**0.156**	**0.077**
NOSEMREL	142	0.07	**-0.148**	**-0.102**	**-0.132**
NODISREL	91	0.05	-0.025	**-0.081**	**-0.062**
misleading discourse connectives⋆	114	0.06	-	-	-
sum(clause level violations)	2,986	1.54	**-0.325**	0.041	-0.016
sum(clause level violations, DNP-REF, PRN+MISSA)	4,305	2.22	**-0.385**	**-0.084**	**-0.122**
sum(all violations)	5,727	2.96	**-0.356**	-0.022	**-0.101**

Further Analysis

- Linear model investigates the relationship of particular errors to readability

Feature	Weight	Feature	Weight
Intercept	3.407	DNP-REF	-0.157
ACR-EXPL	-0.361	OTHRUNGR	-0.155
PRN+MISLA	-0.355	INCLDATE	-0.151
INCOMPLSN	-0.275	INP+REF	-0.067
NOSEMREL	-0.262	NODISREL	-0.046
REDUNDINF	-0.259	FM-EXPL	-0.023
PRN+MISSA	-0.236	SM+EXPL	0.038

- Most significant factors: Missing/Misleading refs, fragments, redundant content, poor coherence
- Total # of errors well-correlated with system ranks

Automatic Evaluation of Linguistic Quality

- Motivation:
  - No focus on linguistic quality b/c no way to tune to it
  - Everyone uses ROUGE b/c you can tune
    - Explicitly tuned in many ML models
- Alternative strategies:
  - Micro: Learn to predict component scores
  - Macro: Learn to predict overall readability score
    - Intuitively: error count (LVQSumm) predicts well, but...
      - Errors manually derived

Micro-Quality Prediction

- (Pitler et al, 2010) via SVM ranking

- Evaluate multiple measures aimed to model LQ
  - General word choice, sequence: Language Models
  - Reference form:
    - Named Entities: modification for 1st mention of PER
    - NP syntax: POS, phrase tags in NPs
  - Local coherence
    - Devices: counts of pron, dem, connectives,...
    - Continuity: adjacency in source, coref w/prev, same, cosine
  - Sentence fluency: features from MT eval
  - Coh-Metrix: set of psycho-ling motivated feats, LSA sim
  - Word coherence: cross-sentence word cooccurence patterns
  - Entity coherence: via Entity-grids (Brown toolkit)

Results

- System level

Feature set	Gram.	Redun.	Ref.	Focus	Struct.
Lang. models	87.6	83.0	91.2	85.2	86.3
Named ent.	78.5	83.6	82.1	74.0	69.6
NP syntax	85.0	83.8	87.0	76.6	79.2
Coh. devices	82.1	79.5	82.7	82.3	83.7
Continuity	88.8	**88.5**	**92.9**	**89.2**	**91.4**
Sent. fluency	**91.7**	78.9	87.6	82.3	84.9
Coh-Metrix	87.2	86.0	88.6	83.9	86.3
Word coh.	81.7	76.0	87.8	81.7	79.0
Entity coh.	90.2	88.1	89.6	85.0	87.1
Meta ranker	**92.9**	87.9	91.9	87.8	90.0

- Summary level

Feature set	Gram.	Redun.	Ref.	Focus	Struct.
Lang. models	66.3	57.6	62.2	60.5	62.5
Named ent.	52.9	54.4	60.0	54.1	52.5
NP Syntax	59.0	50.8	59.1	54.5	55.1
Coh. devices	56.8	54.4	55.2	52.7	53.6
Continuity	61.7	62.5	**69.7**	**65.4**	**70.4**
Sent. fluency	**69.4**	52.5	64.4	61.9	62.6
Coh-Metrix	65.5	**67.6**	67.9	63.0	62.4
Word coh.	54.7	55.5	53.3	53.2	53.7
Entity coh.	61.3	62.0	64.3	64.2	63.6
Meta ranker	**71.0**	**68.6**	**73.1**	**67.4**	**70.7**

Findings

- Overall accuracies quite good
- Systems overall easier to rank than particular input
  - Smoothes variance, larger sample
- Continuity related features best across components
  - Ensemble of ordering, coref, cosine similarity cues
    - Though LSA-based system detects redundancy well
- Specifically tuned fluency scorer works on fluency

Macro-Quality Prediction

- (Lin et al, 2012) Downloadable

- High-level idea:
  - Discourse version of entity grid
    - Columns: entities (same head)
    - Rows: sentences
    - Cell values: PDTB Relation.Arg# tuples
- Variants:
  - Inter-cell sequence frequencies
    - + Additional tuples: {Non--}Explicit.Relation.Arg#
    - + Intra-cell “sequences”

Results

- Very strong correlations w/manual readability score
  - Beats prior predictors

	Initial			Update		
	P	S	K	P	S	K
R-2	0.7524	0.3975	0.2925	0.6580	0.3732	0.2635
R-SU4	0.7840	0.3953	0.2925	0.6716	0.3627	0.2540
BE	0.7171	0.4091	0.2911	0.5455	0.2445	0.1622
4	0.8194	0.4937	0.3658	0.7423	0.4819	0.3612
6	0.7840	0.4070	0.3036	0.6830	0.4263	0.3141
12	0.7944	0.4973	0.3589	0.6443	0.3991	0.3062
18	0.7914	0.4746	0.3510	0.6698	0.3941	0.2856
23	0.7677	0.4341	0.3162	0.7054	0.4223	0.3014
LIN	**0.8556**	**0.6593**	**0.4953**	**0.7850**	**0.6671**	**0.5008**
LIN+C	**0.8612**	**0.6703**	**0.4984**	**0.7879**	**0.6828**	**0.5135**
LIN+E	**0.8619**	**0.6855**	**0.5079**	**0.7928**	**0.6990**	**0.5309**
DICOMER	**0.8666**	**0.7122**	**0.5348**	**0.8100**	**0.7145**	**0.5435**

Referring to People in News Summaries

- Intuition:
  - Referring expressions common source of errors
  - References to people prevalent in news data, summaries
  - Information status constrains realization
  - Targeted rewriting can improve readability
- Approach:
  - Exploit information status distinctions
    - Automatically identified
  - Use to guide rule-based generation of referring expressions

Challenges

- Lack of training data:
  - No summary data labeled for information status

- Readers sensitive to referring expressions
  - Prior work on NP rewriting has shown mixed results
    - Some improvement, some failures

- Relies on potentially errorful coref, other processing

NP Rewrite: very good example

- While the British government defended the arrest, it took no stand on extradition of Pinochet to Spain, leaving it to the courts.

- While the British government defended the arrest in London of former Chilean dictator Augusto Pinochet, it took no stand on extradition of Pinochet to Spain, leaving it to British courts.

NP Rewrite: mixed example

- *Duisenberg* has said growth in the euro area countries next year will be about 2.5 percent, lower than the 3 percent predicted earlier.
- *Wim Duisenberg, the head of the new European Central Bank*, has said growth in the euro area countries next year will be about 2.5 percent, lower than just 1 percent in the euro-zone unemployment predicted earlier.

Information Status

- Build on three key distinctions:
  - Discourse-new vs discourse-old:
    - First mention handling vs others
  - Hearer-new vs hearer-old:
    - Distinguish well-known individuals from others
      - Don't waste space describing well-known individuals
  - Major vs minor character:
    - Salience of the person in the event

Corpus Analysis

- Assess relation between:
  - information status and referring expressions

		Discourse-new	Discourse-old
Name Form	Full name	0.97	0.08
	Surname only	0.02	0.87
	Other (e.g., Britney, JLo)	0.01	0.05
Pre-Modification	Any	0.51	0.21
	None	0.49	0.79
Post-Modification	None	0.60	0.89
	Apposition	0.25	0.04
	Relative clause	0.07	0.03
	Other	0.08	0.04
Any Modification	Some Modification	0.76	0.30
(Either Pre- or Post-)	No Modification	0.24	0.70

Generating Discourse-New/Old

- If discourse-new,
  - If the NP head is a person name,
    - If appears with pre-modifier in text, write as:
      - Longest pre-modifier + full name
    - Else if it appears with an apposition modifier
      - Add that to the reference
  - Else don’t rewrite
- Else use surname only
- Significantly preferred over original forms

Hearer & Salience

- Discourse-new status:
  - Obvious from summary
- How do we establish hearer or major/minor status?
- Categorize based on human summaries (gold)
  - Specifically by their referring expressions:
    - Hearer-old (i.e. familiar)
      - Title/role+surname or unmodified fullname
    - Major:
      - Referred to by name in some human summary of topic
        - 258 major/3926 minor by data

Training & Application

- Trained classifiers to recognize – using features in docset
  - H-New/Old: F-measure: 0.75 on both classes
  - Major/Minor: F: Major: 0.6; Minor: 0.98
    - All significantly better than baseline
- Create rules based on classification to:
  - Use names (only) for major characters (o.w. common)
  - Exclude post-modifiers for hearer-old, tune for new
  - Include titles for initial mentions
  - Include affiliation premodifiers based on hearer/salience

Evaluation

- Created (nearly) deterministic rule set
  - Based on information status classification
  - To rewrite referring expressions in extractive summaries
- Evaluated in paired preference tests over:
  - Original Extractive and Rewritten Summaries
- Where a preference was expressed,
  - Rewritten summaries rated as more coherent
  - Extractive rated as more informative
    - Why? Rewrite rules generally shrink rather than add content

Summary

- Can identify particular correlates of readability scores
- Can automatically predict linguistic quality scores
- Build systems that focus on frequent violations
  - Yield systematic improvements in linguistic quality