The Contribution of Domain-independent Robust Pronominal Anaphora Resolution to Open-Domain Question-Answering

Rebecca Watson, Judita Preiss, Ted Briscoe
Computer Laboratory
University of Cambridge

- OD QA and Anaphora
- Baseline QA System
- Performance on TREC 8 Data
- TREC 8 QA Data Analysis
- Robust Anaphora Resolution
- Contribution to QA
- Conclusions / Further Work

QA and Anaphora

QA Task Definition

TREC: document-based OD QA
TREC 8 QA dataset (top 1000)
TREC 8 Gold standard and eval (MRR)
BUT return full sentence! (approx. 250byte task)

Contribution of Anaphora

What country is the biggest producer of tungsten?: China

> The 15 countries attending the three-day annual market review, which ended yesterday, account for about 90 per cent of world trade in tungsten products. They include China, the biggest producer, which represents over 60 per cent of world trade...

They = The 15 countries
China = a country
The biggest producer = producer of tungsten (products)

How much of a help how often?

Baseline QA System

‘Glue’ = robust Minimal Recursion Semantics (rMRS):

Elementary Predications:
tungsten(x1), **product(x2)**, **ARGN u1 x1**...
Variable sorts: objects, **x**, events, **e**, underspecified, **u**
Variable equality statements: **x1=x2**

LKB + LingERG grammar – parse questions into rMRSs
RASP System – parse top documents into PSTs/GRs/rMRSs

Matching
Match question rMRSs to document sentence rMRSs:

Named entity recognition / classification
Morphological analysis
Expansion of predicates (WordNet)
etc

Weighted sum of (in)directly matched elements of rMRSs

RASP System Outputs

```
(|T/txt-sc1/---|
 (|T/leta_s|
  (|S/s_co_np1|
   (|S/np_vp| |They_PPHS2|
    (|V/np| |include_VV0|
      (|NP/n1_name/-|
          (|N1/n| |China_NP1|))))
   |,_,|
   (|NP/det_n| |the_AT|
    (|N1/ap_n1/-|
      (|AP/a1| (|A1/a| |biggest_JJT|))
     (|N1/n| |producer_NN1|))))
  (|Tacl/comma-e| |,_,|
   (|S/whnp_vp| |which_DDQ|
    (|V/np| |represent+s_VVZ|
     (|NP/ap2_np| (|A1/a| |over_RP|)
      (|NP/plu3|
       (|N1/num2_nms|
          (|NP/num| (|N1/n| 60_MC))
        (|N1/nms_nms| |per_NNU|
         (|N1/n_of| |cent_NNU|
          (|PP/p1|
           (|P1/p_n1| |of_IO|
            (|N1/n1_nm| |world_NN1|
                (|N1/n| |trade_NN1|
                     ))))))))))))))
```

```
GRs:
(ncsubj represent+s_VVZ which_DDQ _)
(dobj represent+s_VVZ cent_NNU _)
(ncsubj include_VV0 They_PPHS2 _)
(dobj include_VV0 China_NP1 _)
(ncmod _ producer_NN1 biggest_JJT)
(detmod _ producer_NN1 the_AT)
(ncmod _ include_VV0 producer_NN1)
(ncmod _ trade_NN1 world_NN1)
(ncmod of_IO cent_NNU trade_NN1)
(ncmod _ cent_NNU per_NNU)
(ncmod _ cent_NNU 60_MC)
(mod _ cent_NNU over_RP)
(cmod _ include_VV0 represent+s_VVZ)

rMRS:
they_rel u2, include_rel u4
ARG1  u4 u2, ARG2  u4 u7
china_rel x6, the_rel x12
biggest_rel x12, producer_rel x12
which_rel x27, represent_rel e29
over_rel e29, 60_rel u33
per_rel x35, cent_rel x37
of_rel e39, ARG2 e39 x41
world_rel x41, trade_rel x50
```

MRR on TREC 8/9 data

TREC 8 (163 questions):

rMRS	0.472
+Morph	0.476
+WordNet+NE	0.484
rMRS+Context	0.619

TREC 9 (10 questions):

rMRS	0.150
+Morph	0.178
+WordNet+NE	0.270
+Context	0.470

'rMRS' = weighted matching
'+Morph' = deriv. morph analysis and matching
'+WordNet+NE' = predicate expansion + NE class mismatch filtering
rMRS+Context = weighted matching returning 5 sentence window

(5 sentences because 98.7% of anaphors have antecedents in previous 2 sentences in this dataset.)

Context matters much more than Morph, NER or WordNet expansion

TREC 8 QA Data Analysis

intraP	0.11
interP	0.04
interD	0.13
contx+	0.14
contx-	0.10

‘intraP’ = intrasentential pronominal anaphora
‘interP’ = intersentential pronominal anaphora
‘interD’ = definite description anaphora (not appos, etc)
‘contx+’ = context inference required (*tungsten*)
‘contx-’ = spurious matches

48% of questions can be answered from the matching sentence

Anaphora resolution is relevant to contextual inference in two thirds of the genuine contextual cases

Robust OD Anaphora Resolution

Lappin & Leass' algorithm, GR-based

Coreference Filters: e.g. Argument Domain Filter

Kim seems to want to see him

```
(ncsubj see_VV0 Kim_NP1 _)
(dobj see_VV0 he_PPHO1 _)
```

```
(arg - X N -)
(arg - X P -)
```

where arg $\in \{ncsubj, dobj, iobj, obj2\}$
X is a variable over predicates
N and P are nominal and pronominal dependents of X

Salience Factors:

There is a Porsche. It is green.

Factor	Weight
Sentence recency	100
Subject emphasis	80
Existential emphasis	70
Accusative emphasis	50
Indirect object/oblique	40
Head noun emphasis	80
Non-adverbial emphasis	50
Parallelism	35
Cataphora	175

Accuracy of LL Reimplementation

	BC	BU	CH	C1	C2
1	60	63	63	63	61
2	51	53	54	55	54
3	70	70	69	67	69
4	67	65	70	64	67
5	55	53	50	52	52
μ	61	61	62	61	61

'BC' = Rasp system parser + GR output
'BU' = Memory-based GR classifier
'CH' = Maxent-inspired PTB parser
'C1' = Collins Model1 PTB parser
'C2' = Collins Model2 PTB parser

Results for 5 annotated portions of BNC (2.4k pronouns)

(No def. descrip. anaphora as is difficult in the (unsupervised) OD context)

No signif. diffs. so RASP-GR+LL = OD pronoun resolution
(as RASP is virtually unlexicalized)

Contribution to QA

RASP-GR+LL resolves 73.2% of pronouns correctly in ‘intraP’ and ‘interP’ TREC 8 5 sentence contexts

36% of errors involve misidentification of the head in the antecedent rather than the antecedent itself (e.g. El in El Nino)

Baseline	0.491
+antecedent	0.510
+direct-subst	0.499
+partial-rMRS	0.483
+full-rMRS	0.459
+context	0.619

‘Baseline’ / ‘+context’ = lower / upper bounds
‘+antecedent’ = manual substit. of antecedent for pronoun
‘+direct-subst’ = auto. addit. of elem. preds. for antec. head
‘+partial-rMRS’ = +elem. preds. linked to antec. head
‘+full-rMRS’ = entire rMRS for sent(s) containing antecedents

In the ‘+antecedent’ condition, 71% of submissions improved but altered MRR for only 10% cases (as intrasent. anaphora was within same submitted sentence).

BUT this would be relevant for 50byte task!

Conclusions / Further Work

- Anaphora resolution is very relevant to OD QA on the TREC 8 dataset

- Probably generalize: questions not based on text content, but scientific texts have more def. descrip. anaphora than newspaper texts

- RASP-GR+LL works well for pronouns in (unsupervised) OD context, but need to extend to def. descripts. and room for improvement: weighted coref. constraints, weight optimization

- Integration of antecedent-related rMRSs from context sentences with matching sentence needs more work as does the rMRS output from the RASP system

Papers, software etc:
http://www.cl.cam.ac.uk/Research/NL/