CS4025: Realisation and Commercial NLG

Realisation

- Third (last) NLG stage
- Generate actual text
- Take care of details of language
  - » Syntactic details
    - – Eg Agreement (the dog *runs* vs the dogs *run*)
  - » Morphological details
    - – Eg, plurals (dog/dogs vs box/boxes)
  - » Presentation details
    - – Eg, fit to 80 column width

Realisation

- Problem: There are lots of finicky details of language which most people developing NLG systems don’t want to worry about
- Solution: Automate this using a realiser

Syntax

- Sentences must obey the rules of English grammar
  - » Specifies which order words should appear in, extra function words, word forms
- Many aspects of grammar are somewhat bizarre

Syntactic Details: Verb Group

- Verb group is the main verb plus helping words (auxiliaries).
- Encodes information in fairly bizarre ways, eg tense
  - » John will watch TV (future – add will)
  - » John watches TV (present - +s form of verb for third-person singular subjects)
  - » John is watching TV (progressive – form of BE verb, plus +ing form of verb)
  - » John watched TV (past – use +ed form of verb)

Verb group

- Negation
  - » Usually add “not” after first word of verb group
    - – John will *not* watch TV
  - » Exception: add “do not” before 1-word VG
    - – Inflections on do, use infinitive form of main verb
    - – John *does not* watch TV

Verb group

- Negation
  - » Usually add “not” after first word of verb group
    - – John will *not* watch TV
  - » Exception: add “do not” before 1-word VG
    - – Inflections on do, use infinitive form of main verb
    - – John *does not* watch TV vs John watch *not* TV
  - » Exception to exception: use first strategy if verb is form of BE
    - – John is *not* happy vs John *does not* be happy

Realiser

- Just tell realiser verb, tense, whether negated, and it will figure out the VG
  - » (watch, future) -> will watch
  - » (watch, past, negated) -> did not watch
  - » Etc
- Similarly automate other “obscure” encodings of information

Other examples

- Adjective ordering
  - » Big red apple vs Red big apple
- Agreement and measurements
  - » Three miles is a long way
  - » Three children are hungry
- Bare infinitives and perception verbs
  - » I see John eat an apple
  - » I see John thinks a lot

Other examples

- Adjective ordering
  - » Big red apple vs Red big apple
- Agreement and measurements
  - » Three miles is a long way
  - » Three children are hungry
- Bare infinitives
  - » I see John eat an apple

Morphology

- Words have different forms
- Nouns have plural
  - » Dog, dogs
- Verbs have base, present, past, present participle, past participle
  - » break, breaks, broke, breaking, broken
- Adjectives have comparative, superlative
  - » Big, bigger, biggest

Formation of variants

- Example: plural
  - » Usually add “s” (*dogs*)
  - » But add “es” if base noun ends in certain letters (*boxes*, *guesses*)
  - » Also change final “y” to “i” (*tries, babies*)
  - » Many special cases
    - – *children* (vs *childs*), *people* (vs *persons*), etc

Realiser

- Calculates variants automatically
  - » (*dog*, plural) -> *dogs*
  - » (*box*, plural) -> *boxes*
  - » (*child*, plural) -> *children*
  - » etc

Morphophonology

- Form of a word depends on following word
  - » an apple vs a banana
- Bigger problem in French than English
- Realiser can automate

Punctuation

- Rules for structures
  - » Sentences have first word capitalised, end in a full stop
    - – My dog ate the meat.
  - » Lists have conjunction (eg, *and*) between last two elements, comma between others
    - – I saw Tom, Sue, Zoe, and Ciaran at the meeting.
  - » Etc
- Realiser can automatically insert appropriate punctuation for a structure

Punctuation

- Rules for structures
  - » Sentences have first word capitalised, end in a full stop
    - – My dog ate the meat.
  - » Lists have conjunction (eg, *and*) between last two elements, comma between others
    - – I saw Tom, Sue, Zoe, and Ciaran at the meeting.
    - – I saw Tom, Sue, Zoe and Ciaran at the meeting.
  - » Etc
- Realiser can automatically insert appropriate punctuation for a structure

Punctuation

- Rules on combinations of punc
  - » Don’t end full stop if sentence already ends in a full stop
    - – He lives in Washington D.C.
    - – He lives in Washington D.C..
  - » Brackets absorb some full stops
    - – John lives in Aberdeen (he used to live in Edinburgh).
    - – John lives in Aberdeen (he used to live in Edinburgh.).
- Again realiser can automate

Pouring

- Usually we insert spaces between tokens, but not always
  - » My dog
  - » Mydog
  - » I saw John, and said hello.
  - » I saw John , and said hello
- Automated by realiser

Pouring

- Often want to insert line breaks to make text fit into a page of given width
  - » Breaks should go between words if possible
    - – Breaks should go between words if poss
      ible
  - » If not possible, break between syllables and add a hyphen
- Realiser automates

Output formatting

- Many possible output formats
  - » Simple text
  - » HTML
  - » MS Word
- Realiser can automatically add appropriate markups for this

Realiser systems

- simplenlg – relatively limited functionality, but well documented, fast, easy to use, tested
- KPML – lots of functionality but poorly documented, buggy, slow
- openccg – somewhere in between
- Many more

(Montreal) French simplenlg

- Vaudry and Lapalme, 2013
- Lots of “silly” rules, like English.
- Eg, negation
  - » il ne parle pas
    - – “he does not speak”
  - » il ne parle plus
    - – “he does not speak anymore”
  - » personne ne parle
    - – “nobody speaks”

Morphophonology

- Many cases in French
  - » le + homme → l'homme
  - » la + honte → la honte
  - » le + beau + homme → le bel homme
  - » à + le → au

Other languages

- Simplenlg for German, Portuguese, etc.
  - » Each language has its “quirks”
  - » German runs words together
    - – Aircraft engine -> Flugzeugtriebwerk
  - » 今天天气很好。->今天|天气|很好。
- Most challenging/different is tribal languages, eg from New Guinea
  - » Allman et al, 2012

Summary

- Realiser automates the finicky details of language
  - » So NLG developer doesn't have to worry about these
  - » One of the advantages of NLG

Commercial NLG

- Arria/Data2text: U Abdn spinout company
  - » Financial reports
  - » Weather forecasts
  - » Explanations of equipment alarms

Others

- Narrative Science - Builds bespoke "automatic narrative generation" systems
  - » Academic roots in computational creativity
- Automated Insights - writes "insightful, personalized reports from your data"
  - » Non-academic roots
- Yseop - "Smart NLG" software that "writes like a human"
  - » Chief scientist, Alain Kaeser did NLG in 1980s

Others

- Lots of small young startups, I lose track of them
  - » OnlyBoth “Discovers New Insights from Data. Writes Them Up in Perfect English. All Automated”
  - » InfoSentience “Developers of the Most Advanced Automated Narrative Generation Software”
  - » Text-on (German) “Aus abstrakten Daten werden so Texte”
- NLG projects at large companies.
  - » INLG 2012 panel - Thomson-Reuters, Agfa
  - » More secretive

Common Themes

- Almost all claim to generate narratives/stories from data
- Financial reporting is most commonly mentioned use
- Companies still quite small
  - » Fewer than 100 employees, compared to 12,000 at Nuance or 400,000 at IBM
  - » But large compared to earlier NLG companies
  - » Also lots of them!

Robojournalism

- Computers write articles for newspapers
  - » Sports, finance, weather
- Lots of media attention
  - » http://www.bbc.co.uk/news/technology-34204052 (many others)

Arria/Data2text history

- 2009: *Data2text* set up
  - » Commercialising research in NLG and data-to-text, esp SumTime and Babytalk
  - » 2 academics, 2 devs, 2 business guys
- 2013: Arria/Data2text goes public
  - » Arria (London-based) bought Data2text
    - – Sales, marketing, corporate, IT architects, project managers to complement techies
  - » Listed on AIM stock market in Dec 2013
  - » About 40 employees

Arria/Data2text now

- Offices in Aberdeen, London, Sydney
  - » Sales teams in London and New York
  - » Growing/hiring (Java devs with NLP/NLG)
  - » 50 developers, 15 commercial
  - » 8 patents, more on the way
  - » www.arria.com

Responsibilities

- Employee (“meet the payroll”)
- Investor (“how is my money doing”)
- Client (“fix this yesterday”)
- End user (“how is my baby doing”)

Success

- NLG/data-to-text essential, but only a small part of overall story
  - » “boring” IT
  - » Support
  - » Change management
  - » Sales and marketing

Questions