NLP Applications and Related Courses Informatics 2A: Lecture 31



School of Informatics University of Edinburgh mlap@inf.ed.ac.uk

1 December 2011

1 Tasks and Applications

2 Sample Applications

- Speech Synthesis
- Information Retrieval
- Machine Translation
- Text Simplification
- Summarization

3 Further Courses



We have discussed several types of natural language processing (NLP) techniques in Inf2a:

- Part of speech tagging: assigning lexical categories to words;
- Parsing: assigning syntactic structure to strings;
- Language Modeling: assigning probability to a string;
- Semantic Analysis: assigning an interpretation to a string.

Here, we'll look at where these techniques can be used in NLP applications to do practical things.



Speech synthesis generates speech from from text. The main components of a Text-to-Speech system:

- Speech database: Recordings of speakers segmented into diphones or larger units;
- Text processing: To normalize the input text (numbers, punctuation, etc.);
- Linguistic processing: To add pronunciation information (phonetic and prosodic) to the text;
- Synthesis: To select speech units from the database and assemble them based on the pronunciation information.

Most of these components are trained on corpus data using probabilistic (machine learning) models.

1/34



Information Retrieval

Information retrieval (IR): Given a query, find all and only documents that are relevant to it, assuming:

- There's a large document collection to search;
- The user has an information need.
- The user's need can be formulated in terms of a query (typically keywords);

Example: search engine

Document collection: web pages. Information need: Pages on a particular topic. Query: user-specified keywords Output: a ranked list of relevant web pages.

Other examples: bibliographic information systems; electronic newspaper archives.

Speech Synthesis

Demos of deployed TTS systems:

- AT&T System, male voice, American English ⁽¹⁾
- AT&T System, female voice, British English
- CereProc System, female voice, Scottish English 🐠
- CereProc System, male voice, American English 🖤

Most commercial systems now use unit selection and can sound very natural in limited domains.

eech Synthesi

6/34

Tasks and Applications Sample Applications Further Courses Summarization

Information Retrieval



Information Retrieval: Benefits from NLP

- Stemming can improve performance by mapping (*walk*, *walks*, *walking*, *walked*) onto the same query term.
- Stop word removal can improve performance by removing words like *a*, *the*, *of*, etc. that don't discriminate between documents.
- NLP can resolve ambiguities:
 - word sense ambiguity (see *tiger* in previous example);
 - part of speech ambiguity: *walk* as noun or verb;
 - syntactic ambiguity: make her duck.
- Logical form of query text and document text might provide an even better basis for matching, but currently not feasible.

Machine Translation

Machine translation (MT) systems turn a text in a source language into a text in a target language (e.g., English to French).

Summarization

To reach human levels of translation requires:

- linguistic knowledge (lexicon, syntax, semantics, pragmatics);
- world knowledge to figure out which how to translate an expression in a given context.
- Harder to formalize than linguistic knowledge.

Current solution: learn from corpora that are translations of each other: machine translation as encryption.



Noisy channel model



- Key idea: treat French output as encrypted English input.
- Large bilingual corpora which reflect similar "noise" can serve for learning how to break the encryption.
- These are available, e.g., in the form of parliamentary proceedings (Canadian Hansard, Hong-Kong Hansard, EU Parliament).



Mathematical formulation of the noisy channel model: to translate a source language sentence F into a target language sentence E compute:

$$\hat{E} = \arg \max_{E} P(E)P(F|E)$$

where P(E) is a language model of the target language and P(F|E) is the translation model.

To compute the translation model, we need to align the bilingual training corpus on the sentence level and on the word level.

9/34

neech Synthesis eech Synthesis Tasks and Applications Sample Applications Tasks and Applications Sample Applications **Machine Translation** Machine Translation Further Courses Further Courses Text Simplification Summarization Machine Translation Machine Translation Step 2: Sentence Alignment Step 2: Sentence Alignment Step 1: Document Alignment Step 1: Document Alignment amor omnia vincit love conquers all Step 3: Word Alignment 13/34 eech Synthesis Tasks and Applications Sample Applications Tasks and Applications Sample Applications Machine Translation Machine Translation Further Courses Further Courses Summarization

Google Translate

French Source

L'année 1866 fut marquée par un événement bizarre, un phénomène inexpliqué et inexplicable que personne n'a sans doute oublié. Sans parler des rumeurs qui agitaient les populations des ports et surexcitaient l'esprit public l'intérieur des continents les gens de mer furent particulièrement émus. Les négociants, armateurs, capitaines de navires, skippers et masters de l'Europe et de l'Amérique, officiers des marines militaires de tous pays, et, après eux, les gouvernements des divers États des deux continents, se préoccupèrent de ce fait au plus haut point.

English Target

Google Translate

The year 1866 was marked by a bizarre event, a phenomenon unexplained and inexplicable that no one has probably forgotten. without about the rumors which agitated the people of the ports and overexcited public opinion within the continents of people sea were particularly moved. Traders, shipowners, captains of vessels, skippers and masters of Europe and America, officers the navies of all countries, and after them, governments the various states of the two continents, were concerned of this fact highest point.

Speech Synthesis Information Retrieval Machine Translation Text Simplification Summarization

Tasks and Applications Sample Applications Further Courses Sumple Applications Further Courses

Google Translate

English Target

The year 1866 was marked by a bizarre event, a phenomenon unexplained and inexplicable that no one has probably forgotten. without about the rumors which agitated the people of the ports and overexcited public opinion within the continents of people sea were particularly moved. Traders, shipowners, captains of vessels, skippers and masters of Europe and America, officers the navies of all countries, and after them, governments the various states of the two continents, were concerned of this fact highest point.

Excerpt from a famous book. Can you guess its title?

16/34

Tasks and Applications Sample Applications Further Courses

Information Retrieva Machine Translation Text Simplification

Human Translation

F. P. Walter

THE YEAR 1866 was marked by a bizarre development, an unexplained and downright inexplicable phenomenon that surely no one has forgotten. Without getting into those rumors that upset civilians in the seaports and deranged the public mind even far inland, it must be said that professional seamen were especially alarmed. Traders, shipowners, captains of vessels, skippers, and master mariners from Europe and America, naval officers from every country, and at their heels the various national governments on these two continents, were all extremely disturbed by the business.

Human Translation

F. P. Walter

THE YEAR 1866 was marked by a bizarre development, an unexplained and downright inexplicable phenomenon that surely no one has forgotten. Without getting into those rumors that upset civilians in the seaports and deranged the public mind even far inland, it must be said that professional seamen were especially alarmed. Traders, shipowners, captains of vessels, skippers, and master mariners from Europe and America, naval officers from every country, and at their heels the various national governments on these two continents, were all extremely disturbed by the business.

17 / 34

Tasks and Applications Sample Applications Further Courses Sumparization

Two Owl Tales

Owls are the order Strigiformes, comprising 200 bird of prey species. Owls

prising 200 bird of prey species. Owls
 hunt mostly small mammals, insects,
 and other birds though some species
 specialize in hunting fish.

An owl is a bird. There are about 200 kinds of owls. Owls' prey may be birds, large insects (such as crickets), small reptiles (such as lizards) or small mammals (such as mice, rats, and rabbits).



Speech Synthesis Information Retrieval Machine Translation Text Simplification Summarization

Two Owl Tales

Owls are the order Strigiformes, comprising 200 bird of prey species. Owls hunt mostly small mammals, insects, and other birds though some species specialize in hunting fish.



An owl is a bird. There are about 200 kinds of owls. Owls' prey may be birds, large insects (such as crickets), small reptiles (such as lizards) or small mammals (such as mice, rats, and rabbits).

Substitute rare words with more familiar words or phrases

Tasks and Applications Sample Applications Further Courses Sample Applications

Two Owl Tales



An owl is a bird. There are about 200 kinds of owls. Owls' prey may be birds, large insects (such as crickets), small reptiles (such as lizards) or small mammals (such as mice, rats, and rabbits).



Remove unnecessary and complicating detail



An owl is a bird. There are about 200 kinds of owls. Owls' prey may be birds, large insects (such as crickets), small reptiles (such as lizards) or small mam-

mals (such as mice, rats, and rabbits).

Simplify deep syntactic structures

Two Owl Tales

-

Tale

Owls are the order Strigiformes, com-

Tasks and Applications Sample Applications

Further Courses

Machine Translation

Text Simplification

prising 200 bird of prey species. Owls hunt mostly small mammals, insects, and other birds though some species specialize in hunting fish.



18/34

- An owl is a bird. There are about 200
 - kinds of owls. Owls' prey may be birds, large insects (such as crickets), small reptiles (such as lizards) or small mammals (such as mice, rats, and rabbits).

Explain unfamiliar words or concepts

Tasks and ApplicationsSpeech SynthesisSample ApplicationsInformation RetrievalFurther CoursesText SimplificationSummarizationSummarization

Two Owl Tales

Owls are the order Strigiformes, comprising 200 bird of prey species. Owls

prising 200 bird of prey species. Owls
 hunt mostly small mammals, insects, and other birds though some species specialize in hunting fish.

18/34

Tasks and Applications Sample Applications Machine Translation Text Simplification Further Courses

Why simplify?

1993 US National Adult Literacy Survey (grades 1-5)

neech Synthesis



Tasks and Applications Sample Applications Machine Translation Further Courses Text Simplification

How to simplify?



Speech Synthesis Tasks and Applications Sample Applications Text Simplification Further Courses Summarization Why simplify?

Percentage of the adult population for each literacy grade



How to simplify?

Simple English Wikipedia is a sort of "spin-off" of Wikipedia.

- Treat SimpleEW edits as instances of simplifications?
- But many edits aren't simplifications.
- Only consider revisions accompanied by "simpl" comments



ns Information Retrieva ns Machine Translation ses Text Simplification Summarization

neech Synthesis

Synchronous Grammars

Synchronous grammars are a way of simultaneously generating pairs of recursively related strings.

- Originally invented for programming language compilation
- Generalization of context-free grammar formalism to simultaneously produce strings in two languages.
- Have been used extensively in **syntax-based SMT:** inversion transduction grammar (ITG; Wu 1997), head transducers (Alshawi et al., 2000), hierarchical phrase-based translation (Chiang, 2007), synchronous tree substitution grammar (STSG; Eisner, 2003), quasi-synchronous grammar (GSG; Smith and Eisner, 2006).

Idea: Use the synchronous grammar formalism to learn simplification paraphrases.











VP

2











A summarization system turns a source document (or multiple documents) into a summary that contains its key points:



Example: search engines often return a summary with each hit.

Speech Synthesis Information Retrieval Machine Translation Text Simplification Summarization

Summarization

A typical summarization system has the following components:

Tasks and Applications Sample Applications

Further Courses

peech Synthesis

Machine Translation

Summarization

- sentence extraction: decides which sentences of the source document to retain in the summary;
- sentence compression: takes the extracted sentences and removes "non-essential" material (e.g., adjuncts);
- sentence combination: combines extracted sentences into longer ones.

Each of these components can benefit from NLP techniques (e.g., POS tagging, parsing).

Machine Translation

Summarization

26 / 34

Tasks and Applications Sample Applications Further Courses

Summarization

Extracted Sentences

S1: Saudi Arabia on Tuesday decided to sign the nuclear weapons non-proliferation treaty, a strong indication it will not seek nuclear warheads for intermediate-range missiles it recently acquired from China.
S2: The official Saudi Press Agency reported that King Fahd made the decision during a Cabinet meeting in Riyadh, the Saudi capital. S3: The meeting was called in response to a recommendation by Prince Saud al-Faisal, the Saudi foreign minister, that the kingdom sign the international treaty against the spread of nuclear arms.

Summarization

Source Text

S1: Saudi Arabia on Tuesday decided to sign the nuclear weapons non-proliferation treaty, a strong indication it will not seek nuclear warheads for intermediate-range missiles it recently acquired from China.
S2: The official Saudi Press Agency reported that King Fahd made the decision during a Cabinet meeting in Riyadh, the Saudi capital. S3: The meeting was called in response to a recommendation by Prince Saud al-Faisal, the Saudi foreign minister, that the kingdom sign the international treaty against the spread of nuclear arms. S4: An account of the Cabinet discussions and decisions at the meeting, which ended before dawn, was issued by Information Minister Ali al-Shaer and distributed by the agency. The agency, monitored in Bahrain, did not elaborate. S5: It appeared the timing of the decision was designed primarily to reassure the United States that the kingdom will not try to arm its CSS-2 missiles with nuclear warheads. The decision also was viewed as an attempt to blunt Israel's allegations that the missiles constituted a threat to its safety.

27 / 34

Tasks and Applications Sample Applications Further Courses Sumplification

Summarization

Reduced Sentences

S1: Saudi Arabia on Tuesday decided to sign the nuclear weapons non-proliferation treaty, a strong indication it will not seek nuclear warheads for intermediate-range missiles it recently acquired from China.
S2: The official Saudi Press Agency reported that King Fahd made the decision during a Cabinet meeting in Riyadh, the Saudi capital. S3: The meeting was called in response to a recommendation by Prince Saud al-Faisal, the Saudi foreign minister, that the kingdom sign the international treaty against the spread of nuclear arms.

Combined Sentences

S1: Saudi Arabia decided to sign the nuclear weapons non-proliferation treaty, a strong indication it will not seek nuclear warheads for intermediate-range missiles. S2+S3: The official Saudi Press Agency reported that King Fahd made the decision during a Cabinet meeting, called in response to a recommendation by Prince al-Faisal.

Speech Synthesis Information Retrieva Machine Translation Text Simplification Summarization

Machine Translation

Summarization

Caption Generation

Caption Generation

Michelle Obama fever hits the UK

In the UK on her first visit as first lady, Michelle Obama seems to be making just as big an impact. She has attracted as much interest and column inches as her husband on this London trip; creating

a buzz with her dazzling outfits, her own schedule of events and her own fanbase. Outside Buckingham Palace, as crowds gathered in anticipation of the Obamas' arrival, Mrs Obama's star appeal was apparent.



30 / 34

Michelle Obama fever hits the UK

In the UK on her first visit as first lady, Michelle Obama seems to be making just as big an impact. She has attracted as much interest and column inches as her husband on this London trip; creating



a buzz with her dazzling outfits, her own schedule of events and her own fanbase. Outside Buckingham Palace, as crowds gathered in anticipation of the Obamas' arrival, Mrs Obama's star appeal was apparent.

30 / 34

Tasks and Applications Sample Applications Further Courses

Caption Generation

Michelle Obama fever hits the UK

In the UK on her first visit as first lady, Michelle Obama seems to be making just as big an impact. She has attracted as much interest and column inches as her husband on this London trip; creating

a buzz with her dazzling outfits, her own schedule of events and her own fanbase. Outside Buckingham Palace, as crowds gathered in anticipation of the Obamas' arrival, Mrs Obama's star appeal was apparent.



Queen, Obama, Michelle, Buckingham, London, UK

It is reported that the Queen asked to stay in touch with Mrs Obama





Given what you now know about Inf2a:

- You'd like to know more about formal languages
- 2 You'd like to know more about NLP
- You'd like to know more about both
- It was a good course but you did not find the material exciting
- You'd rather switch to drama than do anything remotely related

Tasks and Applications Sample Applications Further Courses Samplification

The Ultimate Clicker Question II

How do I do well on the exam?

- Read the slides and do the reading
- 2 Read the slides and go over past exams
- Read the slides, do the reading, and can answer tutorial questions with ease
- O nothing, just chance it
- 6 Revise the Pumping Lemma

Summarization

The Ultimate Clicker Question II

How do I do well on the exam?

- Read the slides and do the reading
- 2 Read the slides and go over past exams
- Read the slides, do the reading, and can answer tutorial questions with ease
- O nothing, just chance it
- **6** Revise the Pumping Lemma

Tip: if we spent more than one lecture on a topic, we consider it important, and chances are it will figure in the exam; ditto for topics figuring in tutorials.

32 / 34

Tasks and Applications Sample Applications Further Courses Courses related to NLP

2nd Year

- Informatics 2B (S2): Algorithms and Pattern Recognition
- Informatics 2D (S2): Modern Artificial Intelligence

3rd Year

• Foundations of Natural Language Processing (S2): continues where Inf2a left off (more NLP algorithms, techniques, tasks)

4th Year

- Speech Processing (S1), Speech Synthesis (S2)
- Text Technologies (S1)
- Machine Translation (S2), Natural Language Understanding (S2), Natural Language Generation (S2)



Courses related to Formal Languages

3rd Year

- Computability and Intractability (S1)
- Compiling Techniques (S2)
- Language Semantics and Implementation (S2)

4th Year

- Advances in Programming Languages (S1)
- Compiler Optimization (S2)
- Computational Complexity (S2)