VRIJE UNIVERSITEIT BRUSSEL
FACULTEIT WETENSCHAPPEN
DEPARTEMENT COMPUTERWETENSCHAPPEN

EXAM

NATURAL LANGUAGE PROCESSING

Monday 19 June 2017
10:00 to 12:00

Examiner: Dr. K. Heylen

INSTRUCTIONS TO CANDIDATES
The exam counts for 50% of your final grade
Questions can be answered in any order.

Make sure to put your name on each sheet of paper you hand in at the end and to sign the list of attendance.

Questions

1. **N-grams**, **multi-word expressions**, and **collocations** are defined in NLP as 3 different types of word combinations. Give an informal definition of each and highlight the properties that differentiate between them. **[5 points]**

2. Write a regular expression to find "irregular" words. Regular words are defined here as strings that are delineated by a whitespace character on both ends, that start with a lowercase or uppercase letter, followed by 1 or more lowercase letters with no other character types allowed ("letter" is here defined as an alphabetic ASCII character, disregarding letters with diacritics such as accents). The "irregular" words that your regular expression should match, are also delineated by whitespace characters and they also begin with a lowercase or uppercase letter, but unlike regular words, the characters that follow (minimally 1) contain at least one uppercase letter OR a non-alphabetic character (e.g. integers, underscores, symbols, hyphens etc.) Examples of irregular words would be: **ParTy, b3have, a$$ets, re_mind, r.s.v.p., note-book** **[5 points]**

3. Explain the difference between the use of a Finite State Transducer as a recognizer and as a translator. Which of these 2 usage types is applied in morphological parsing? Explain why this usage type is best suited for the output that a morphological parser should generate. **[5 points]**

4. The tables below show the raw unigram and bigram counts for 8 words (out of V = 1446) in a corpus of 9332 consecutive sentences. Given these counts, describe the steps and calculations you need to do to compute the bigram-based estimate for the probability of the sentence "I want Chinese food". If you do not have a calculator handy, you do not have to do the calculations by hand, but you should give the relevant formulae with the counts filled out for the example sentence. **[5 points]**

UNI-GRAMS	i	want	to	eat	chinese	food	lunch	spend		
	2533	927	2417	746	158	1093	341	278		
BIGRAMS	<s>	i	want	to	eat	chinese	food	lunch	spend	</s>
<s>	0	633	2	0	5	22	123	23	12	0
i	0	5	827	0	9	0	0	0	2	2
want	0	2	0	608	1	6	6	5	1	22
to	0	2	0	4	686	2	0	6	211	156
eat	0	0	0	2	0	16	2	42	0	98
chinese	0	1	0	0	0	0	82	1	0	12
food	0	15	0	15	0	1	4	0	0	743
lunch	0	2	0	0	0	0	1	0	0	122
spend	0	1	0	1	0	0	0	0	0	23
</s>	9331	0	0	0	0	0	0	0	0	0
a sample of unigram and bigram counts for 8 words (out of V = 1446), plus the sentence start and end symbol, in a toy corpus with 9332 consecutive sentences										

5. POS-tagging approaches that are based on Hidden Markov Models make two important simplifying assumptions that can be formally expressed by the two approximations below that hold between different types of probabilities for words (w) and POS-tags (t) in a word sequence. Give an informal description of the simplifying assumptions expressed by in the approximation below. Explain how these two simplifying assumptions are also directly related to the two types of probabilities that can link states in a Hidden Markov Model. Make sure that you give the names for these two types of HMM probabilities and that you describe which type of probability can link between which types of state in an HMM for POS-tagging. **[10 points]**

Simplifying Assumption 1	$P(w_1^n \mid t_1^n) \approx \prod_{i=1}^{n} P(w_i \mid t_i)$
Simplifying Assumption 2	$P(t_1^n) \approx \prod_{i=1}^{n} P(t_i \mid t_{i-1})$

6. The equation below represents the value function that statistical machine translation systems using a noisy channel approach try to maximize in order to find the best possible translation of a sentence from the source language F into the target language E. The two probabilities in the equation are the statistical operationalisation of two graded properties that have traditionally been central to evaluating translation quality, but that are often difficult to reconcile. Which are these two properties and why is there often a trade-off between them? Additionally, these two probabilities in the equation also represent two statistical models that are trained separately and then combined in the value function used for finding optimal translations. Which are the 2 statistical models represented by P(F|E) and by P(E) respectively? Finally, briefly discuss why it might be at first sight counterintuitive to have the conditional probability P(F|E) instead of P(E|F) in the equation but explain why this is nonetheless correct. **[5 points]**

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