On Frequency-based Approaches to Learning Stopwords and the Reliability of Existing Resources

A Study on Italian Language

Stefano Ferilli, Floriana Esposito
Dipartimento di Informatica - Università di Bari, Bari, Italy

stefano.ferilli@uniba.it

IRCDL 2018
14TH ITALIAN RESEARCH CONFERENCE ON DIGITAL LIBRARIES
UDINE - JANUARY 25-26, 2018

Overview

- Introduction & Motivation
- Current Landscape & Objectives
- Proposed Approach
- Quantitative & Qualitative Evaluation
- Conclusions & Future Work

Introduction

- Most DL content is text
- NLP techniques of utmost importance for the proper management of DLs
- Based on language-specific linguistic resources
- Might be unavailable for many languages
- Manual compilation costly, time-consuming and error-prone
- Desirable to learn these resources automatically
- Often no prior knowledge about the language

Background

- BLA-BLA tool
  - Broad-spectrum Language Analysis-Based Learning Application
  - Fully automatic learning of linguistic resources from plain text(s) in a given language
    - Language identification
    - Stopword removal
    - Term normalization
    - Concept extraction
  - Works on very small corpora
  - General approaches applicable to any language
    - Terms from other languages = Noise

Focus

- Stopwords
  - Terms that are not necessary to understand the topic and content of a document
    - Appear often and pervasively
    - Have the same likelihood of occurring in documents not relevant to a query as in those relevant to the query [IR]
  - By definition, can be safely ignored by NLP techniques that work at the lexical level
    - Removal task simply carried out by look-up in a pre-determined list of words

Basics

- Stopword removal
  - Early step in NLP pipeline
  - May affect the performance of subsequent steps
- Stopword Lists
  - Function words
    - Terms associated to invariant Parts-of-Speech of the language (usually articles, pronouns and prepositions)
      - Requires prior knowledge about the grammar of the language
  - Frequent terms
    - Domain-specific terms in domain-specific applications

Current Landscape

- Past approaches to learning stopword lists
  - Vector Space Model
    - Based on Porter's stemmer
      - Language-dependent
      - Requires language-specific tools/resources
  - Purely frequency-based approaches
    - Deal with specific languages
      - English, French
    - Million words corpora
    - Manual adjustment of the learned list of stopwords
- Benchmark stopword lists (freely) available

Objectives

- Experimental study on frequency behavior of words
    - Assessment of quality and reliability of existing resources
- Technique for automatic support to stopword list compilation
    - Language-independence
        - May be used for non-widespread languages (e.g., dialects)
    - Small training data
        - Quality of the results for increasingly larger data
        - Different ages and styles

Proposed Approach

- Very simple
  - Extract multiset W of words in the corpus
    - V “vocabulary” (the set of different words in W)
  - Compute relative frequency of each word $w \in W$
    - $f(w) = |w|/|W|$
      - Ratio of number of occurrences of the word over the overall number of word occurrences in the text(s)
  - Rank members of V by decreasing frequency
  - Consider the set S of all words $v \in V$ for which $f(v) \geq f'$ for a frequency threshold f'
  - Check for stopwords in S

Experimental Setting

- Word
  - Sequence of alphabetic characters only, delimited by blank spaces or punctuation
    - Apostrophe joining two words was considered as well
  - Formally defined by the linear expression pattern:
    - b P { W' }* W P b
      - b the blank symbol
      - ' the apostrophe
      - P = { .|,|;|:|?|!|"|' \}* (possibly empty) sequence of punctuation marks
      - W = { a|b|...|z }^+ word (hypothesizing a latin alphabet)

Experimental Setting

- Italian language
  - Has attracted some attention from the NLP community
    - Existing stopword lists may serve as a golden standard
  - Less studied than English
    - Existing resources may be less refined
  - More complex structure than English
    - Experimental results should apply to most other languages, as well
- Small training corpus
  - Stress the proposed approach
    - In large corpora the frequency of real stopwords is clearly predominant
    - For some languages (e.g., dialects) only very few written texts are available

Training Corpus

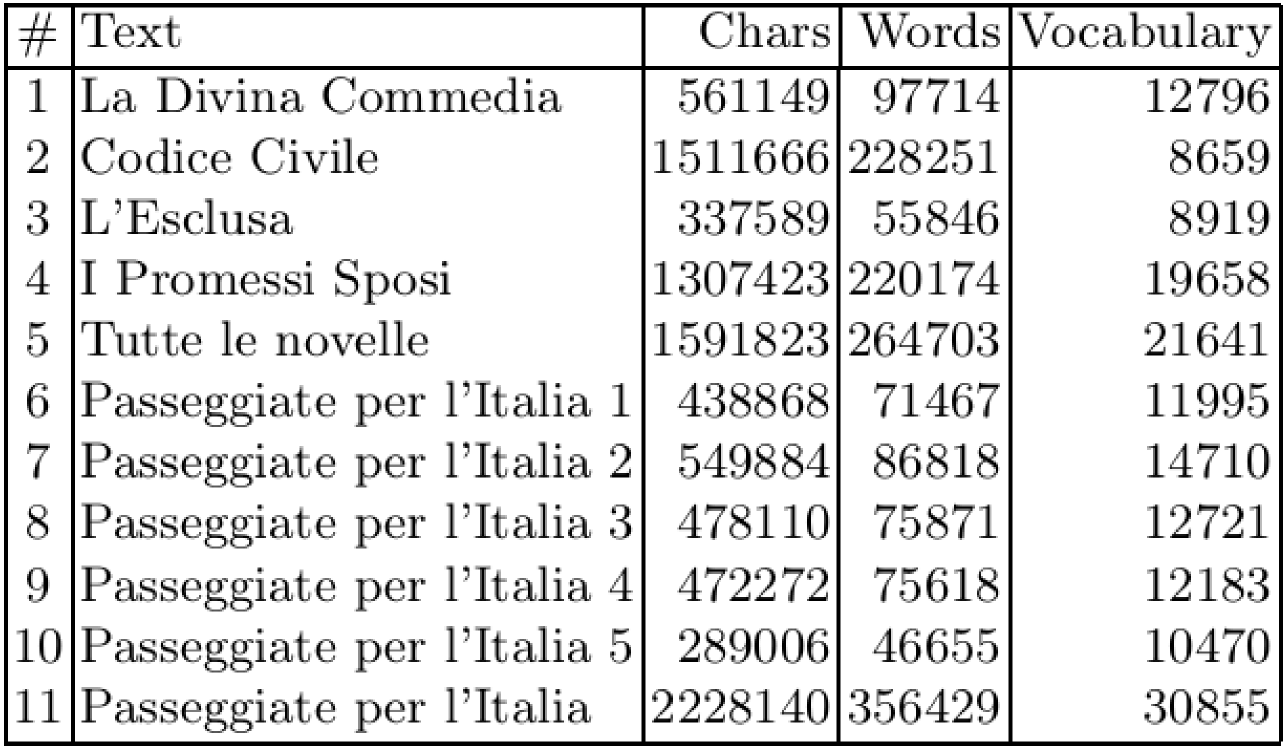
- 10 texts
  - Project Gutenberg and Liber Liber repositories
    - Make freely available many well-known texts from the literature of several languages
    - Obtained by applying OCR to books, and so they contain spelling errors spread through the text
      - Allows us to test our approach on noisy data, which are what one may expect to have in real-world settings
  - Wide range of styles
    - 2 “technical”
      - Poetry, Legal
    - 3 “non-technical”
      - Novels, Stories, Travel accounts

Training Corpus

- Texts
  - La Divina Commedia, poem, XIV century
  - Codice Civile, technical text, XX century
  - L'Esclusa, novel, 2nd half of XIX century
  - I Promessi Sposi, novel, 1st half of XIX century
  - Tutte le novelle, collection of stories, XIX-XX centuries
  - Passeggiate per l'Italia, description of travels, XIX century
- Golden Standard
  - Stopword list provided by Snowball
    - Well-known tool exploited by many systems
    - 279 stopwords (complete form)

Training Corpus

- Statistics
  - Length (number of characters and of words)
    - Approximate (counted by a text editor)
  - Linguistic variety (number of words in `Vocabulary')
    - Exact (computed by the pre-processing step)



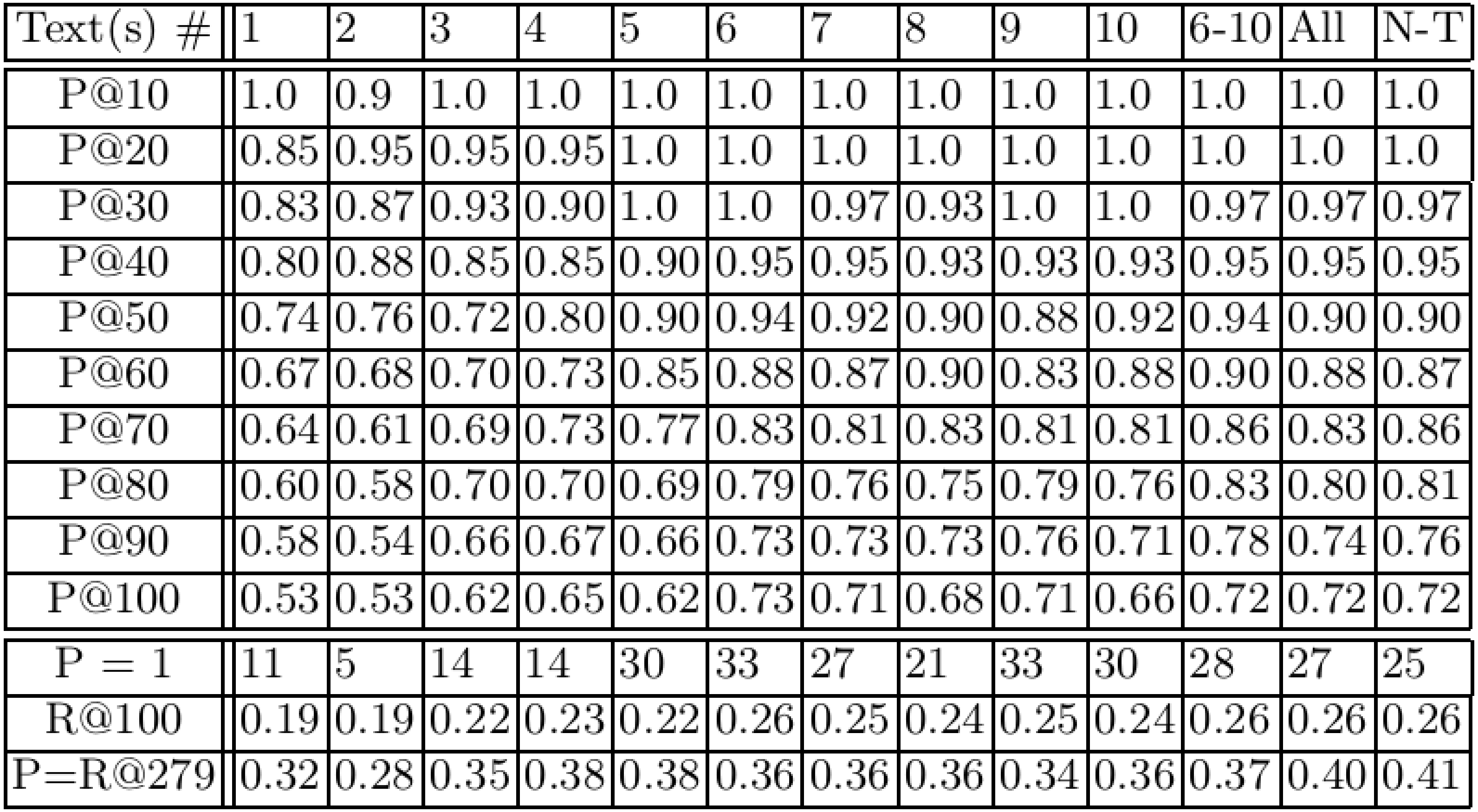
#	Text	Chars	Words	Vocabulary
1	La Divina Commedia	561149	97714	12796
2	Codice Civile	1511666	228251	8659
3	L'Esclusa	337589	55846	8919
4	I Promessi Sposi	1307423	220174	19658
5	Tutte le novelle	1591823	264703	21641
6	Passeggiate per l'Italia 1	438868	71467	11995
7	Passeggiate per l'Italia 2	549884	86818	14710
8	Passeggiate per l'Italia 3	478110	75871	12721
9	Passeggiate per l'Italia 4	472272	75618	12183
10	Passeggiate per l'Italia 5	289006	46655	10470
11	Passeggiate per l'Italia	2228140	356429	30855

Performance Evaluation

- Measures
  - P@*n* : Precision of the top *n* items in the ranking
    - % of items that are also in the golden standard
      - *n* = 100 delimits a ‘safety region’ including (almost) only stopwords
  - P = 1 : maximum position in the ranking at which 100% precision is preserved
    - Indication of how reliable is the top of the ranking
  - R@100 : recall at position 100
    - Compared to P@100 gives an idea of how much of the golden standard is still missing at that point in the list
    - Maximum recall reachable @100 is 100/279 = 0.36
  - P=R@279 : performance at position 279
    - At this position, precision and recall take the same value (P=R)

Experimental Results

- Single text (#) and Relevant aggregates of texts
  - ‘6-10’ = the whole ‘Passeggiate per l'Italia’
  - ‘All’ = the whole set of texts
  - ‘N-T’ = non-technical texts only



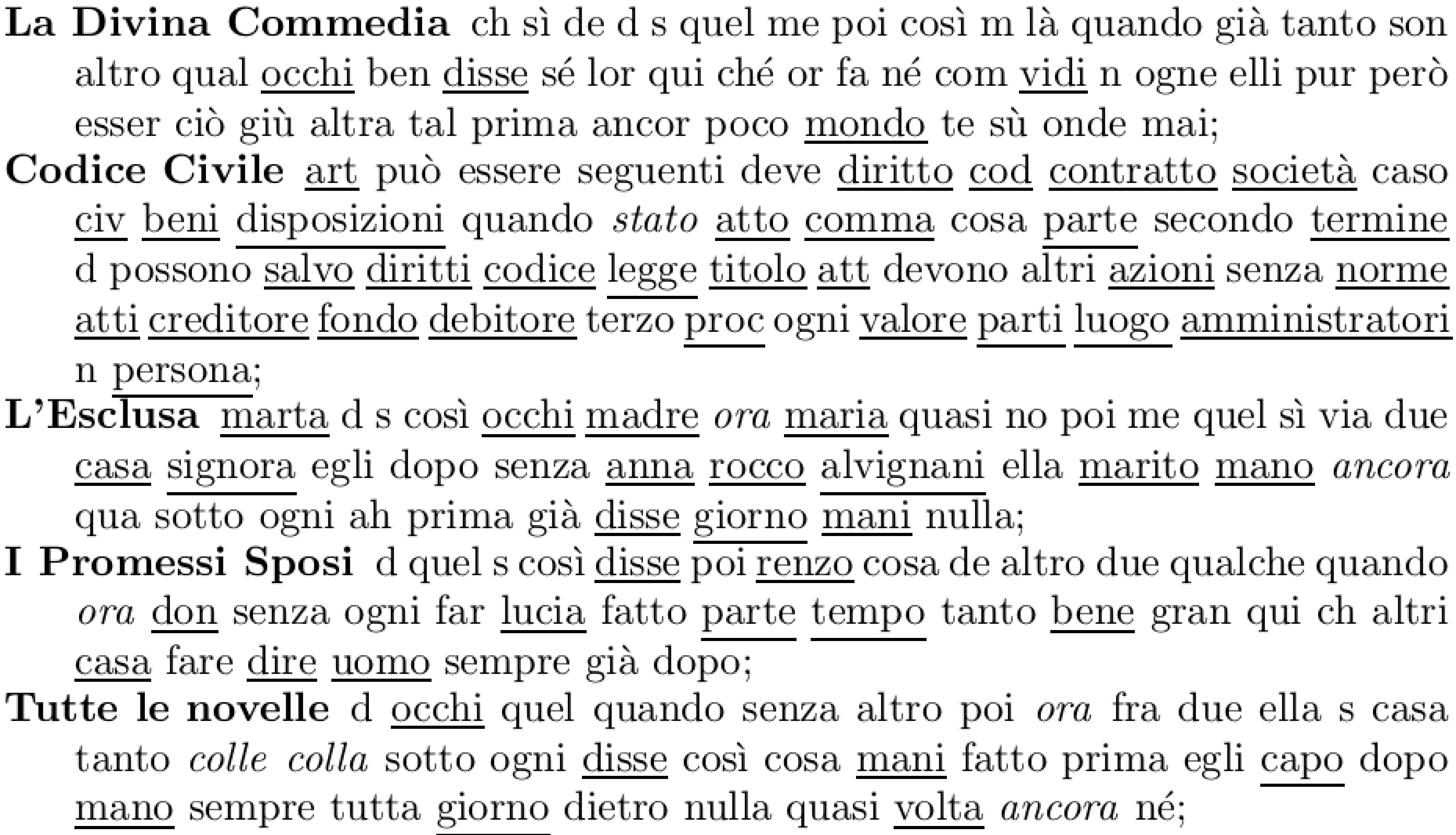
Text(s) #	1	2	3	4	5	6	7	8	9	10	6-10	All	N-T
P@10	1.0	0.9	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
P@20	0.85	0.95	0.95	0.95	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
P@30	0.83	0.87	0.93	0.90	1.0	1.0	0.97	0.93	1.0	1.0	0.97	0.97	0.97
P@40	0.80	0.88	0.85	0.85	0.90	0.95	0.95	0.93	0.93	0.93	0.95	0.95	0.95
P@50	0.74	0.76	0.72	0.80	0.90	0.94	0.92	0.90	0.88	0.92	0.94	0.90	0.90
P@60	0.67	0.68	0.70	0.73	0.85	0.88	0.87	0.90	0.83	0.88	0.90	0.88	0.87
P@70	0.64	0.61	0.69	0.73	0.77	0.83	0.81	0.83	0.81	0.81	0.86	0.83	0.86
P@80	0.60	0.58	0.70	0.70	0.69	0.79	0.76	0.75	0.79	0.76	0.83	0.80	0.81
P@90	0.58	0.54	0.66	0.67	0.66	0.73	0.73	0.73	0.76	0.71	0.78	0.74	0.76
P@100	0.53	0.53	0.62	0.65	0.62	0.73	0.71	0.68	0.71	0.66	0.72	0.72	0.72
P = 1	11	5	14	14	30	33	27	21	33	30	28	27	25
R@100	0.19	0.19	0.22	0.23	0.22	0.26	0.25	0.24	0.25	0.24	0.26	0.26	0.26
P=R@279	0.32	0.28	0.35	0.38	0.38	0.36	0.36	0.36	0.34	0.36	0.37	0.40	0.41

Discussion about Performance

- More related to writing style than to length
  - Makes sense but partly unexpected
  - Colloquial styles more usefulthan technical ones
    - Best on journalistic (‘Passeggiate per l'Italia’)
    - Still quite high on stories (‘Tutte le novelle’)
    - novels come immediately after
    - Lower on the texts written using more particular styles
      - ‘Codice Civile’ (technical) and ‘La Divina Commedia’ (poetry)
  - Using many texts improves performance (expected)
    - improvement not outstanding compared to some single texts, especially for the upper part of the ranking, a smoother decay in performance is clearly visible, as confirmed by the neat increase in performance @279.

Detailed Results

- Single texts



La Divina Commedia ch sì de d s quel me poi così m là quando già tanto son altro qual occhi ben disse sé lor qui ché or fa né com vidi n ogne elli pur però esser ciò giù altra tal prima ancor poco mondo te sù onde mai;

Codice Civile art può essere seguenti deve diritto cod contratto società caso civ beni disposizioni quando *stato* atto comma cosa parte secondo termine d possono salvo diritti codice legge titolo att devono altri azioni senza norme atti creditore fondo debitore terzo proc ogni valore parti luogo amministratori n persona;

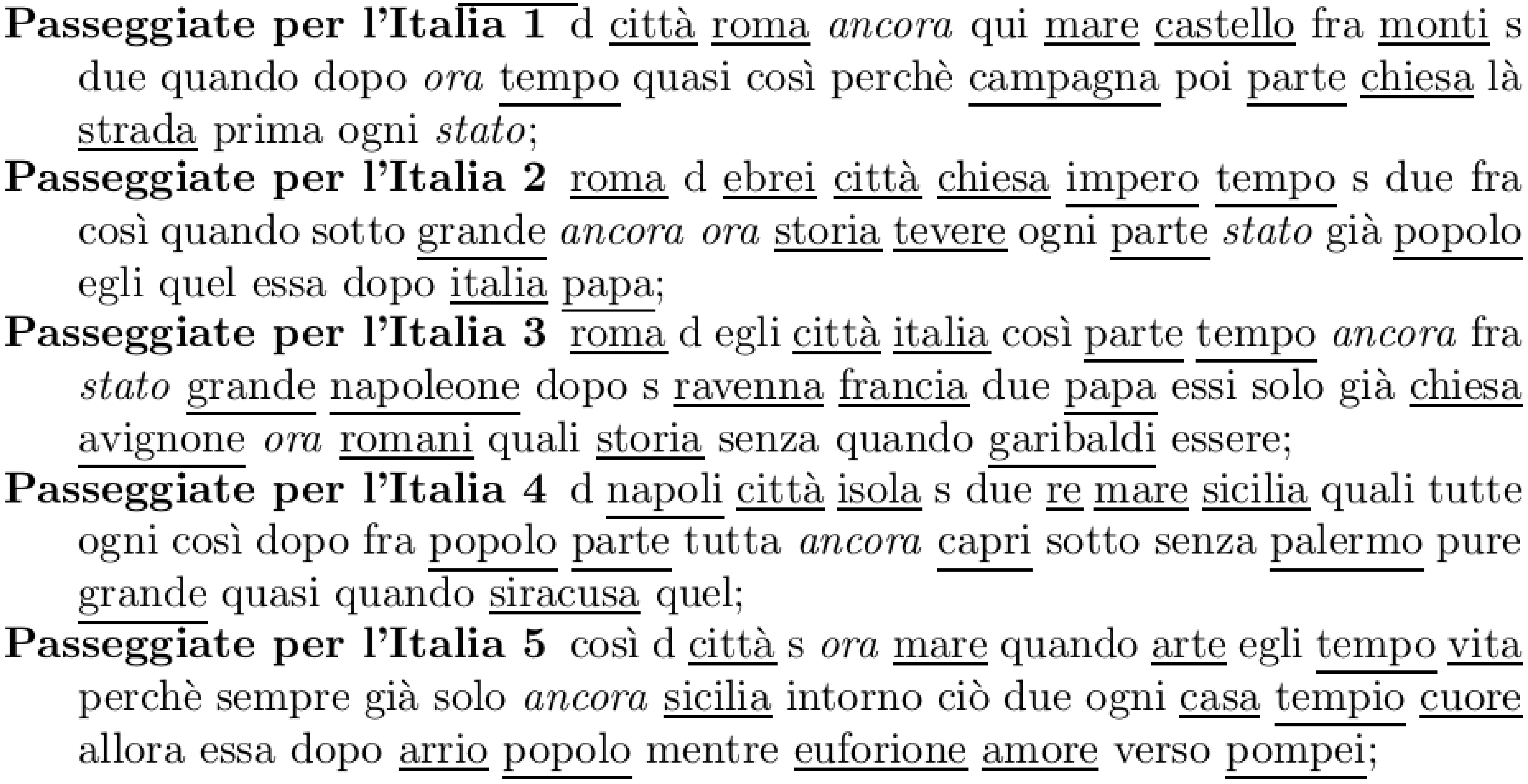
L'Esclusa marta d s così occhi madre *ora* maria quasi no poi me quel sì via due casa signora egli dopo senza anna rocco alvignani ella marito mano *ancora* qua sotto ogni ah prima già disse giorno mani nulla;

I Promessi Sposi d quel s così disse poi renzo cosa de altro due qualche quando *ora* don senza ogni far lucia fatto parte tempo tanto bene gran qui ch altri casa fare dire uomo sempre già dopo;

Tutte le novelle d occhi quel quando senza altro poi *ora* fra due ella s casa tanto *colle colla* sotto ogni disse così cosa mani fatto prima egli capo dopo mano sempre tutta giorno dietro nulla quasi volta *ancora* né;

Detailed Results

- Single texts



Passeggiate per l'Italia 1 d città roma *ancora* qui mare castello fra monti s due quando dopo *ora* tempo quasi così perchè campagna poi parte chiesa là strada prima ogni *stato*;

Passeggiate per l'Italia 2 roma d ebrei città chiesa impero tempo s due fra così quando sotto grande *ancora* *ora* storia tevere ogni parte *stato* già popolo egli quel essa dopo italia papa;

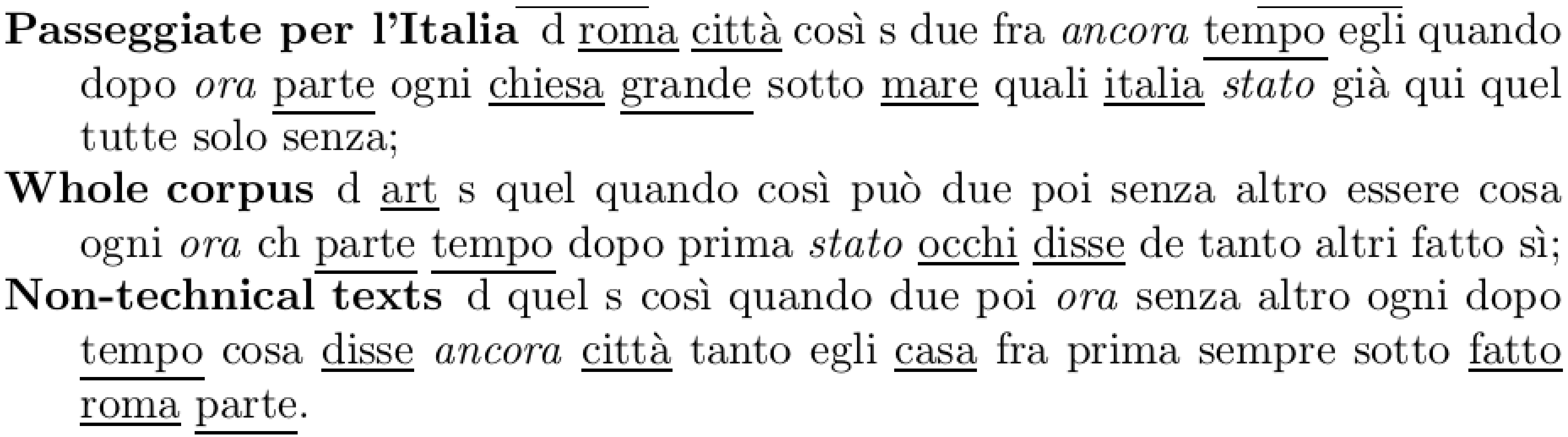
Passeggiate per l'Italia 3 roma d egli città italia così parte tempo *ancora* fra *stato* grande napoleone dopo s ravenna francia due papa essi solo già chiesa avignone *ora* romani quali storia senza quando garibaldi essere;

Passeggiate per l'Italia 4 d napoli città isola s due re mare sicilia quali tutte ogni così dopo fra popolo parte tutta *ancora* capri sotto senza palermo pure grande quasi quando siracusa quel;

Passeggiate per l'Italia 5 così d città s *ora* mare quando arte egli tempo vita perchè sempre già solo *ancora* sicilia intorno ciò due ogni casa tempio cuore allora essa dopo arrio popolo mentre euforione amore verso pompei;

Detailed Results

- Relevant aggregates of texts



Passeggiate per l'Italia d roma città così s due fra *ancora* tempo egli quando dopo *ora* parte ogni chiesa grande sotto mare quali italia *stato* già qui quel tutte solo senza;

Whole corpus d art s quel quando così può due poi senza altro essere cosa ogni *ora* ch parte tempo dopo prima *stato* occhi disse de tanto altri fatto sì;

Non-technical texts d quel s così quando due poi *ora* senza altro ogni dopo tempo cosa disse *ancora* città tanto egli casa fra prima sempre sotto fatto roma parte.

Evaluation of the Golden Standard

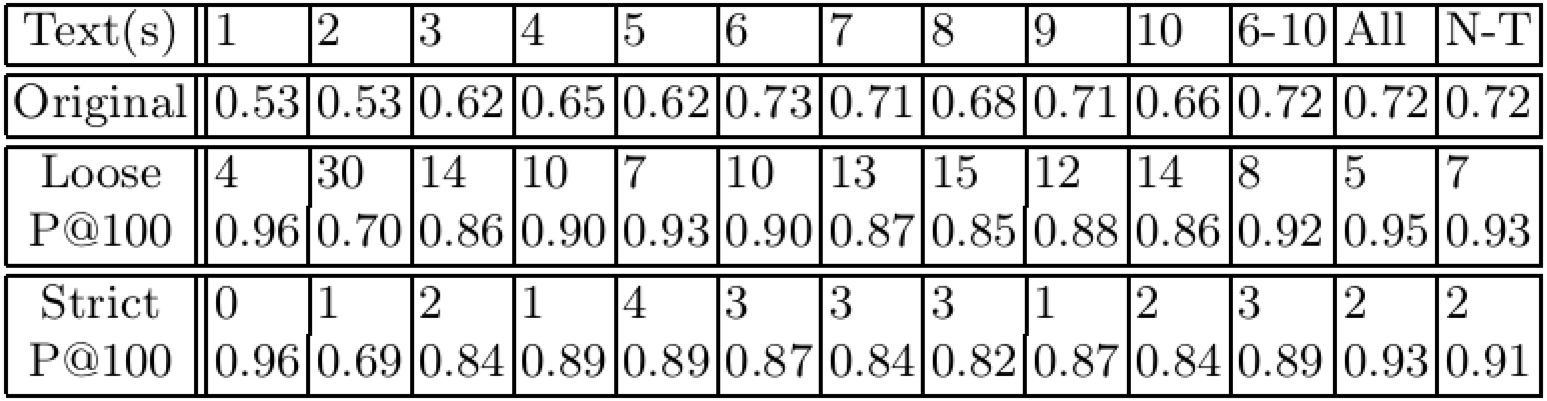
- Missing stopwords
  - Many words in the list that we would safely consider as stopwords are not in the golden standard
  - The absence of some is really strange
    - Many pronouns and generic adverbs (but other similar pronouns or generic adverbs are)
    - Essere (but many inflected form are)
    - Fra (but ‘tra’ is)
    - Etc.
- Conclusions
  - Albeit Italian is a language that received significant attention, the available resources are not reliable

Beyond the Golden Standard

- Re-compute performance
  - Stopwords = all words that do not have a definite meaning by themselves
    - Articles, pronouns, conjunctions and prepositions
    - Some adverbs and some verbs (e.g., modal verbs)
  - Some words are ambiguous
    - Stopwords or not depending on interpretation
      - *stato*: noun (non-stopword) or past participle (stopword)?
      - *colla*: ‘glue’ (non-stopword) or contraction of `con la' (stopword)?
      - *colle*: ‘hill’ (non-stopword) or contraction of `con le' (stopword)?
      - *ancora*: ‘anchor’ (non-stopword) or `still, again' (stopword)?
      - *ora*: ‘hour’ (non-stopword) or `now' (stopword)?
      - ...

Adjusted Evaluation

- Measure
  - Count & P@100
- 2 settings
  - Strict: does not consider ambiguous terms as stopwords
  - Loose: considers ambiguous terms as stopwords



Text(s)	1	2	3	4	5	6	7	8	9	10	6-10	All	N-T
Original	0.53	0.53	0.62	0.65	0.62	0.73	0.71	0.68	0.71	0.66	0.72	0.72	0.72
Loose P@100	4 0.96	30 0.70	14 0.86	10 0.90	7 0.93	10 0.90	13 0.87	15 0.85	12 0.88	14 0.86	8 0.92	5 0.95	7 0.93
Strict P@100	0 0.96	1 0.69	2 0.84	1 0.89	4 0.89	3 0.87	3 0.84	3 0.82	1 0.87	2 0.84	3 0.89	2 0.93	2 0.91

Further Considerations

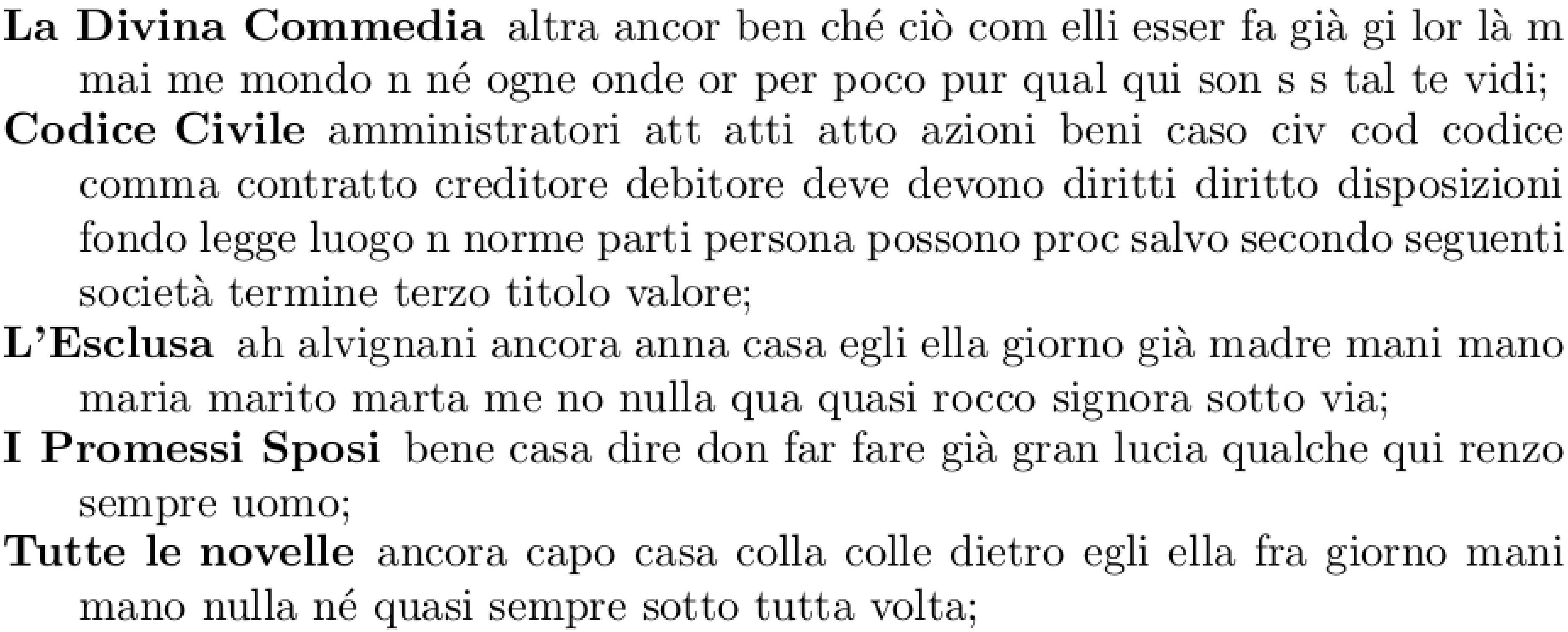
- Texts perspective:
  - Poem includes many stopwords in truncated form
    - Considering them as correct stopwords would make this text the best one, instead of the worst
  - Technical text includes many specific words
    - Misses many stopwords, because they are seldom used in the specific domain
      - Still the worst, even after the corrections are applied
      - Together with non-technical texts improves performance
  - After corrections, novels become the best-performing non-technical single texts
    - Same ‘strict’ performance as the ‘journalistic’ text(s)
    - Even better than them in the loose setting

Further Considerations

- Terms/Stopwords perspective:
  - Using sets of texts wrong terms are pushed towards the end of the list
    - Larger corpora improve the quality of the results
  - Some terms might be considered as stopwords even if missing in the golden standard
    - Terms appearing in all lists
      - E.g., *d*, a truncation of preposition *di*
    - Terms appearing in the majority of lists
      - quando, così, dopo, due, ogni, ora, ancora, già, parte, quel, senza
    - Terms appearing in almost all lists
      - E.g., ‘ora’ and ‘ancora’

Further Perspectives

- Consider terms in the ranking that are not stopwords



La Divina Commedia altra ancor ben ché ciò com elli esser fa già gi lor là m mai me mondo n né ogne onde or per poco pur qual qui son s s tal te vidi;

Codice Civile amministratori att atti atto azioni beni caso civ cod codice comma contratto creditore debitore deve devono diritti diritto disposizioni fondo legge luogo n norme parti persona possono proc salvo secondo seguenti società termine terzo titolo valore;

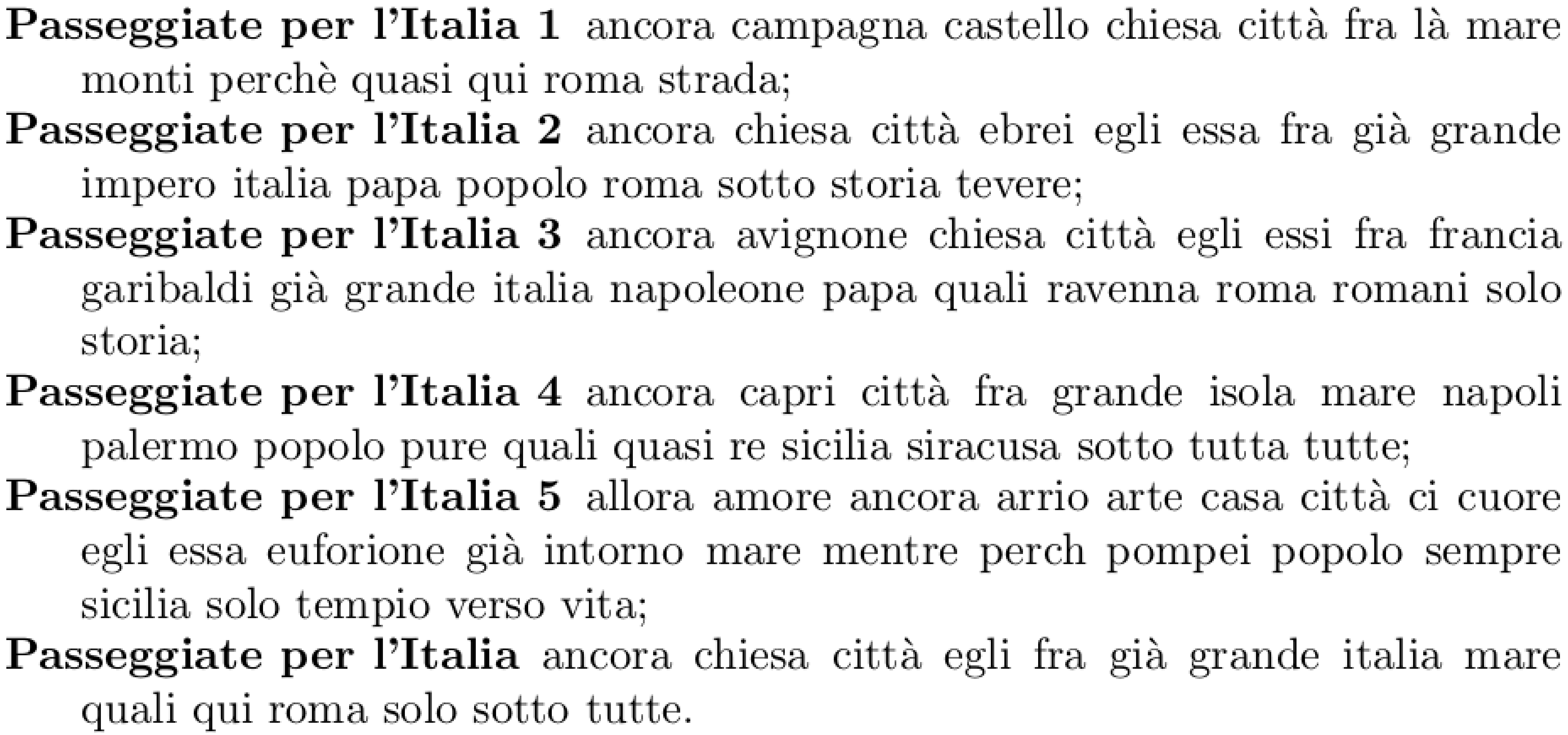
L'Esclusa ah alvignani ancora anna casa egli ella giorno già madre mani mano maria marito marta me no nulla qua quasi rocco signora sotto via;

I Promessi Sposi bene casa dire don far fare già gran lucia qualche qui renzo sempre uomo;

Tutte le novelle ancora capo casa colla colle dietro egli ella fra giorno mani mano nulla né quasi sempre sotto tutta volta;

Further Perspectives

- Consider terms in the ranking that are not stopwords



Passeggiate per l'Italia 1 ancora campagna castello chiesa città fra là mare monti perchè quasi qui roma strada;

Passeggiate per l'Italia 2 ancora chiesa città ebrei egli essa fra già grande impero italia papa popolo roma sotto storia tevere;

Passeggiate per l'Italia 3 ancora avignone chiesa città egli essi fra francia garibaldi già grande italia napoleone papa quali ravenna roma romani solo storia;

Passeggiate per l'Italia 4 ancora capri città fra grande isola mare napoli palermo popolo pure quali quasi re sicilia siracusa sotto tutta tutte;

Passeggiate per l'Italia 5 allora amore ancora arrio arte casa città ci cuore egli essa euforione già intorno mare mentre perch pompei popolo sempre sicilia solo tempio verso vita;

Passeggiate per l'Italia ancora chiesa città egli fra già grande italia mare quali qui roma solo sotto tutte.

Considerations

- Non-stopwords might act as keywords
  - Reading them one may infer that
    - ‘La divina commedia’ is a poem due to the presence of many truncated words
    - ‘Codice Civile’ is about regulations and agreements among people
    - ‘I Promessi Sposi’ and ‘L'esclusa’ are novels, due to the presence of persons' nouns (their main characters are clearly highlighted)
      - In particular, L'Esclusa is about family relationships
    - Passeggiate per l'Italia is about geography/landscape, history/politics and art
      - First three volumes concern Rome
      - Last two concern the Reign of the Two Sicilies

Proposal

- Extending BLABLA
  - Improving stopword extraction feature
  - Adding a keyword extraction feature
- Given a set of texts
  - – Extract candidate stopwords using the frequency-based approach
    - One text: domain-specific terms in the list might be considered as domain-specific stopwords, according to the literature
  - – Compare the stopwords extracted from the complete corpus to the stopwords extracted from the single texts
    - May be used both to identify real stopwords and to extract keywords describing the specific content of the single texts

Conclusions

- Studied the behavior of frequent words in single texts and (small) corpora
- Proposed, based on the study, a methodology to automatically learn stopword lists from texts
  - Also relevant keywords may be extracted with a little extension of the proposed approach
- Preliminary experimental results
  - show that the extracted stopwords and keywords are appropriate
  - pointed out deficiencies of standard resources available in the literature

Future Work

- Define an approach to determine the threshold at which distinguishing stopwords from non-stopwords
- Study of the behavior on larger and more varied corpora
- Indirect evaluation of the quality of results through the performance of high-level NLP tasks based on the learned resources