

D-13: Task-based evaluation report: document classification

Version 1
14 July 2008



Project: STE05039

<http://www.let.vu.nl/onderzoek/projectsites/cornetto/start.htm>

Funded by the Stevin framework (www.taalunieversum.org/stevin)

Authors:

Agata Cybulska, Elton Glaser and Piek Vossen, Irion Technologies

Table of contents

1. Introduction.....	3
2. Word combinations in the Cornetto database	4
3. Evaluation.....	6
3.1 Test set evaluation.....	6
3.2. Evaluation of the indexes	8
4. Conclusions	12

1. Introduction

Document classification is one of the core technologies of Irion Technologies. The classification uses a traditional vector space model to associate topics with training documents. Unseen documents are compared to the training documents to assign classes with scores. The indexes are based on word distributions but the text is processed using natural language tools (NLP tools) to provide optimal indexes. The NLP tools involve normalization and compound splitting. The software package includes a benchmark environment to derive precision and recall figures for a classifier that is trained.

A proper evaluation of the Cornetto database would involve extensive semantic processing that includes named-entity-recognition, word-sense-disambiguation and semantic role labeling of text. With the framework of the Cornetto project, we only had limited resources to carry out an evaluation. We therefore decided to use the database in a very simple and basic way to detect useful combinations of words, regardless of their meaning. The disadvantage of this evaluation is that it does not show the full capacity of the semantic database; the advantage is that integration in the current system is easy and positive results could lead to fast integration in the current product without slowing down the processing time for training and classification. A proper evaluation would require a more extensive project and funding.

The standard Irion system decomposes Dutch compounds into the most salient elements and extends the index with the compound elements in addition to the compound itself. For the evaluation of the Cornetto database, we tried the opposite. If two words in context have a special (semantic) relation with each other, then we extend the index with the word combination, where the words in the combination are normalized and stored in alphabetic order.

This document is structured in the following way: In the next section, we explain how we acquired the word combination from the Cornetto database. Section 3 explains the evaluation framework and Section 4 explains the evaluation results.

2. Word combinations in the Cornetto database

The Cornetto database consists of 3 major files:

- cdb_lu.xml: the lexical units
- cdb_synset.xml: the synsets in which the lexical units combine
- cdb_cid.xml: special identifiers that combine lexical units and synsets and point back to the original sources of data

To derive the word combinations we only used the lexical units and the synsets. For each word in the lexical unit database (a form unit within a lemma), we extracted the following combination words:

- the **form-spelling** attribute field
- the **sem_resume** contents in each **semantics_noun**, **semantics_verb**, or **semantics_adj** block
- the **canonicalform** and **textualform** in each **examples** block

From the synset data, we extracted for each form:

- In the **cdb_syn.xml** file
- In each **wn_internal_relations** block
- the **target-previewtext** contents from each **wn_internal_relations** block
- the **preview_text** string from each synonym

We further skipped stop words and words shorter than 3 characters and we removed duplicates. The resulting lexicon has 60,262 records for unique forms with 396,348 word combination words. Table-1 shows a fragment of the records that were derived:

Table 1: Fragment of multiword lexicon derived from Cornetto

helmstok	lange roer korte zitten
schimmelkleurig	grijswit
roulatiesysteem	afpraak
huisnummer	vermelden numero postcode vergeet
grammaticaliteit	precisie correctheid
hyper(a)emie	accumulatie opeenstapeling ophoping verdichting concentratie stoornis agglomeratie opeenhoping bloed cumulatie
ratificatie	homologatie maastricht confirmatie verdrag bekrachtiging bevestiging bezegeling
attent	stuurde vond anders altijd doelman gaat kaartje mis erg heel iets voortdurend maken voorkomend oplettend gebaar vrouw
praatziek	erg pratend
gravenhuis	geslacht stam familie maagschap
pestblaar	etterzak abces etterbui
determinatie	schepping opwekking
afzijdig	vinden houden nederland meedoend blijven
climax	ogenblik zenit punt tel stijlmiddel hoogtepunt dramatische toewerken suprême seconde moment bereiken stijlfiguur culminatiepunt orgasme minuut
boetedoening	onthechting straf daad asceet vasten afstraffing bloeden iets ascese bidden ascetisme bestraffing versterving boeten quadrageen
grutterij	molen

patrones	jeugd schutspatroom maria patroonheilige patroon schutsheilige beschermheilige
bedroefd	gezicht iets verdrietig
borrel	drinken aanbieden ski lusten feest geven gereformeerde après ophebben stevige aangeklede scheelt slok party pakken nemen teveel
marktsituatie	toestand bedoening
attest	medisch
belastingvlucht	belastingfraude belastingontduiking
polyfaag	gedierte beest dier
ruimtevaart	verkeer luchtvaart
multiraciaal	commissie rassen gemêleerd samenleving gemengd multiraciale
zijdgeweer	schermdegen floret dagge steekwapen schermen fleuret rapier stokdegen dolk degensliker houwdegen
clinch	liggen
stiltegebied	natuurgebied landschapspark

Note that we abstracted from the word meaning and the part of speech of a word. Differentiating these only makes sense if we are able to assign these with high confidence before we do the expansion. As explained in the introduction, we do have sufficient resources within the project to develop high-quality pos-detection and word-sense-disambiguation before we do the multiword matching.

We also think such pre-processing may not be necessary, since the word combinations are already disambiguating and do not leave much room for multiple interpretations and ambiguity for classification.

3. Evaluation

For the evaluation, we used a training set of Dutch text that was built by Irion Technologies for IPTC classification. IPTC stands for the International Press Thesaurus Committee, who maintains an international thesaurus for news articles. This classification scheme is currently used by various Dutch publishers.

For the evaluation we considered the following different methods:

1. The standard Irion system without using the Cornetto data
2. Indexes and classified documents extended with all bigram combinations of content words
3. Indexes and classified documents extended with word combinations from the Cornetto database if they occur in a context window of 10 words
4. Method 2 and 3 are combined where method 2 precedes 3

Classification systems at Irion Technologies are called monks, so from now on we use the word monk for a classification system. The following four monks have been built for the evaluation of the Cornetto database:

1. No-context Monk
2. Lexicon Monk
3. Bigram Monk
4. Both Monk.

Two types of assessment have been completed to evaluate the four monks. The first assessment is a test set evaluation (see 3.1) and the second is an analysis of the indexes of the monks (see 3.2).

3.1 Test set evaluation

Each of the four monks has been evaluated with a test set of 40 XML files that has been used before for the evaluation of the IPTC Monk. The test folder evaluation was performed with six different thresholds: 60%, 65%, 70%, 75%, 80% and 85%.

The tables below show three kinds of evaluation scores for each of the monks tested with different thresholds. Recall, precision and coverage of the monks have been measured. The recall is the average number of correct classes assigned from all the correct classes for each document. Precision is the average number of correct classes assigned out of all classes assigned. Coverage is the proportion documents that have scores above the threshold. The F-score is simply the average of recall and precision.

Table 2: Results of the test set evaluation for the No-context Monk.

Monk	No-context			
Score Threshold	Recall in %	Precision in %	Coverage in %	F-score in %
60%	76.2	19.3	100	47.75
65%	71.2	25.6	100	48.4
70%	70.4	34.5	100	52.45
75%	65.4	42.4	100	53.9
80%	58.5	52.5	97.5	55.5
85%	54.8	63.9	77.5	59.35

Table 3: Results of the test set evaluation for the Bigram Monk.

Monk	Bigram			
Score Threshold	Recall in %	Precision in %	Coverage in %	F-score in %
60%	74.1	20.6	100	47.35
65%	71.6	25.4	100	48.5
70%	70.4	36.6	100	53.5
75%	66.6	43.7	100	55.15
80%	59.6	52.8	95	56.2
85%	54.5	62.8	82.5	58.65

Table 4: Results of the test set evaluation for the Lexicon Monk.

Monk	Lexicon			
Score Threshold	Recall in %	Precision in %	Coverage in %	F-score in %
60%	76.2	20.6	100	48.4
65%	72.5	25.6	100	49.05
70%	71.6	34.1	100	52.85
75%	65.4	43.5	100	54.45
80%	59.4	52.3	92.5	55.85
85%	59.9	64.4	75	62.15

Table 5: Results of the test set evaluation for the Both Monk.

Monk	Both			
Score Threshold	Recall in %	Precision in %	Coverage in %	F-score in %
60%	75	20	100	47.5
65%	73.7	25.7	100	49.7
70%	70.4	34.6	100	52.5
75%	65.4	45.1	100	55.25
80%	55.4	52.4	92.5	53.9
85%	55	61.9	75	58.45

The differences across the monks are low. All four monks have the highest recall and precision scores by the threshold of 85%. The evaluation showed that the Lexicon Monk classified texts with a slightly higher recall and precision score than the other monks. The best F- score of the Lexicon Monk is 62.15% by the coverage of 75%. This is an increase of 2.8 on the baseline system. Recall is 4.5 higher and precision is 0.5 higher than the baseline.

3.2. Evaluation of the indexes

The indexes of the four monks were also the topic of analysis. The table below shows the relevant statistics.

Table 6: Statistics on the indexes

Statistics Monks	Number of multiwords	Number of single words	Total number of lines
No-context	7 ¹	78505	78,512
Bigram	285928	78506	364,434
Lexicon	19379	78505	97,884
Both	312067	78506	390,573

Clearly, the indexes of the multiword monks are much bigger. The bigram monk has 4.6 times more index items and the both monk even 4.97 times more. The lexicon monk has only 1.2 times more index items. This means that the classification is much faster, while low threshold results are slightly lower and high-threshold results are better. Likewise, we can say that the expansion of the index through the lexicon is more effective than random bigram expansion. It could be that a more directed expansion by more carefully selecting the combination words would make the expansion more effective.

The next two tables show some of the multiwords from the lexicon index and the bigram index.

Table 7: Normalized multiword index using Cornetto

aal*paling	aanbellen*bij	aanbidder*heeft	aanbieden*aanbieder	aanbieden*euro
aanbieden*excuses	aanbieden*geven	aanbieden*iets	aanbieden*koop	aanbieden*laten
aanbieden*petitie	aanbieden*tegen	aanbieder*bedrijf	aanbieder*zorg	aanbieding*geldt
aanblik*bij	aanbod*bieden	aanbod*breed	aanbod*gebruik	aanbod*interessant
aanbod*jaar	aanbod*krijgen	aanbod*maken	aanbod*vraag	aanbod*wet

¹ There seem to be some language entities in the index of the No-context Monk that are being treated as multiword units:

noc*nsf

*tijdens

*de

**

noc*nsf-limiet

*beste

noc*nsf-eis.

Task-based evaluation report: document classification

aandacht*afleiden	aandacht*alle	aandacht*bekijk en	aandacht*besteden	aandacht*brengen
aandacht*houden	aandacht*iets	aandacht*krijgen	aandacht*richten	aandacht*ruim
aandacht*schenken	aandacht*speciaal	aandacht*trekke n	aandacht*vergroot	aandacht*verleggen
aandacht*vestigen	aandacht*vragen	aandachtig*iets	aandeel*aandelen	aandeel*bijdragen
aandeel*dalen	aandeel*effect	aandeel*fonds	aandeel*hoger	aandeel*iets
aandeel*lager	aandeel*marktpositi e	aandeel*onderne ming	aandeel*procent	aandeel*staat
aandeel*stijgen	aandeel*waarde	aandeelhouder*b elangrijkste	aandeelhouder*enige	aandeelhouders*belang
aandeelhouders*div idend	aandeelhouders*uit keren	aandeelhouders* vergadering	aandeelhoudersvergad ering*algemene	aandeelhoudersvergaderi ng*buitengewone
aandelen*beleggen	aandelen*cumulatie f	aandelen*uitgifte	aandelen*uitwisseling	aandelen*vermogen
aandienen*gaan	aandoen*iets	aandoen*licht	aandoen*naam	aandoening*chronisch
aandoening*erfelijk e	aandoening*ziekte	aandoeningen*se ksueel	aandragen*oplossinge n	aandringen*erop
aaneengesloten*geb ied	aaneengesloten*geh eel	aaneengesloten*l aatste	aangaan*bondgenootsc hap	aangaan*concurrentie
aangaan*concurrent ieslag	aangaan*confrontati e	aangaan*debat	aangaan*dialogoog	aangaan*fusie
aangaan*samenwer kingsverband	aangaan*strijd	aangaan*volgens	aangeboden*personeel	aangedaan*vrouw
aangedaan*was	aangehouden*verda chte	aangehouden*ve rhoor	aangelegd*maken	aangelegenheden*bemoei en
aangelegenheid*dur e	aangenaam*verblijf	aangenaam*verr ast	aangenomen*eigen	aangenomen*werk
aangepast*gemaakt	aangepaste*versie	aangeschreven*g oed	aangeschreven*hoog	aangeschreven*staan
aangeslagen*zwaar	aangesloten*netwer k	aangetoond*verb and	aangeven*bij	aangeven*duiden
aangeven*moesten	aangeven*wilde	aangewezen*gev al	aangewezen*iets	aangewezen*januari
aangewezen*meest	aangewezen*persoo n	aangewezen*voll edig	aangewezen*waren	aangezet*dik
aangezet*zwaar	aangifte*afpersing	aangifte*diefstal	aangifte*valse	aangifte*vermissing
aangrijpen*gelegen heid	aanhang*trouwe	aanhanger*carav an	aanhanger*vrachtwage n	aanhangwagen*auto
aanhouden*goed	aanhouden*verdach ten	aanhoudende*dr oogte	aanjagen*vuur	aankijken*iets
aankijken*moeten	aankijken*ogen	aankijken*tegen	aanklacht*indienen	aanklacht*luidt
aanklacht*tegen	aanklager*openbaar	aanklager*openb are	aankloppen*bij	aanknopen*banden
aanknopen*relatie	aanknopingspunten *onvoldoende	aankomen*beste mning	aankomen*bij	aankomen*kwam
aankomend*bedrijf	aankomend*talent	aankomst*bij	aankomst*tijd	aankoop*bij
aankoop*kopen	aanleg*beginnen	aanleg*bouw	aanleg*nieuwe	aanleg*ondertunneling
aanleg*plan	aanleg*weg	aanleg*wegen	aanleggen*schip	aanleggen*tegen
aanleggen*tuin	aanleggen*weg	aanleiding*direct	aanleiding*geven	aanleiding*onmiddellijk
aanleiding*zien	aanleren*iets	aanloop*lange	aanloop*nemen	aanlopen*bij
aanlopen*feiten	aanmaning*betalen	aanmerkelijk*risi co	aanmerking*huursubs idie	aanmerking*nemen
aanmerking*vergoe ding	aanmerking*vluchte lingenstatus	aanmeten*laten	aannemelijk*iets	aannemelijk*maken
aannemen*geloof	aannemen*houding	aannemen*krijge n	aannemen*ontslaan	aannemen*personeel
aannemen*smeerge ld	aannemen*steekpen ningen	aannemen*werk	aannemer*opdracht	aanpak*bepaalde
aanpak*harde	aanpak*plan	aanpak*repressie	aanpak*volgen	aanpakken*criminaliteit

		f		
aanpakken*flink	aanpakken*gaan	aanpakken*gaat	aanpakken*goed	aanpakken*hard
aanpakken*werken	aanpakken*wordt	aanpassen*drastisch	aanpassen*omstandigheden	aanpassen*wordt
aanpassing*grondwet	aanpassing*maken	aanplant*nieuwe	aanraking*justitie	aanranding*verkrachting
aanrichten*schade	aanrichten*vernielingen	aanrijden*auto	aanrijding*plaats	aanrijding*veroorzaken
aanroepen*god	aanschaf*auto	aanschaf*kopen	aanschaf*overname	aanschouwen*bewondering
aanschouwen*vol	aanslaan*bij	aanslag*aanval	aanslag*belasting	aanslag*brengen
aanslag*definitieve	aanslag*doel	aanslag*doelwit	aanslag*gericht	aanslag*jaar
aanslag*leven	aanslag*manier	aanslag*minuut	aanslag*plegen	aanslag*portemonnee
aanslag*slachtoffer	aanslag*verantwoordelijkheid	aanslag*voorlopige	aanslepen*brengen	aansluiten*bij
aansluiten*meer	aansluiten*naadloos	aansluiten*rij	aansluiting*lijn	aansluiting*toetreding
aansluiting*zoeken	aanspannen*zaak	aansporen*harder	aansporen*oproepen	aanspraak*maken
aansprakelijk*hoofdelijk	aansprakelijk*stellen	aanspreekbaar*was	aanspreken*gaan	aanstaande*moeder
aanstaande*woensdag	aanstellen*curator	aanstellen*geraakte	aanstellen*kiezen	aanstelling*augustus
aanstelling*krijgen	aanstelling*loopt	aanstelling*traineer	aanstelling*vaste	aansturen*apparatuur

Table 8: Normalized multiword index using bigrams

aanbaksel*goed	aanbellen*avond	aanbellen*bij	aanbellen*straat	aanbesteden*al
aanbesteden*bedrijf	aanbesteden*europes	aanbesteden*fictief	aanbesteden*gaan	aanbesteden*hebben
aanbesteden*mede	aanbesteden*onderhand	aanbesteden*openbaar	aanbesteden*plaat	aanbesteden*strategisch
aanbesteden*stroom	aanbesteden*worden	aanbestedingsbeleid*limburg	aanbestedingsdeskund*p	aanbestedingsdeskund*volgens
aanbestedingsprocedure*europes	aanbestedingsprocedure*melde	aanbestedingsrecht*hoogleraar	aanbestedingsrecht*w	aanbestedingsregel*eigen
aanbestedingsregel*europes	aanbestedingsregel*hebben	aanbestedingsspecialist*bij	aanbetaling*do	aanbetaling*garanderen
aanbetaling*tulp	aanbevelen*aantal	aanbevelen*al	aanbevelen*alom	aanbevelen*ander
aanbevelen*anti	aanbevelen*bij	aanbevelen*do	aanbevelen*enkel	aanbevelen*hebben
aanbevelen*komen	aanbevelen*moet	aanbevelen*nagenoeg	aanbevelen*tour	aanbevelen*vanwege
aanbevelen*worden	aanbevelen*zit	aanbevelen*zullen	aanbidden*arabie	aanbidde n*destijds
aanbidden*fixer	aanbidden*keizer	aanbidden*kunne	aanbidden*leiden	aanbidden*mars
aanbidden*ster	aanbidden*veel	aanbidden*worden	aanbieden*aegon	aanbieden*al
aanbieden*aldus	aanbieden*baan	aanbieden*begeleidende	aanbieden*benefietwedstrijd	aanbieden*beurs
aanbieden*bij	aanbieden*bouwfonds	aanbieden*buitenland	aanbieden*coke	aanbieden*contract
aanbieden*cursus	aanbieden*datingsit	aanbieden*die	aanbieden*dinsdagmiddag	aanbieden*drenthe
aanbieden*dvd	aanbieden*effectendienst	aanbieden*eretitel	aanbieden*eu	aanbieden*even
aanbieden*evenement	aanbieden*excuus	aanbieden*feestdagenprijs	aanbieden*gaan	aanbieden*golfclub
aanbieden*gratis	aanbieden*haaientangd	aanbieden*haar	aanbieden*hebben	aanbieden*helikopter
aanbieden*hotelkamer	aanbieden*hulp	aanbieden*iets	aanbieden*industrie	aanbieden*internet
aanbieden*jaar	aanbieden*juni	aanbieden*kaar	aanbieden*kabel	aanbieden*kamer

Task-based evaluation report: document classification

aanbieden*koerier	aanbieden*komen	aanbieden*koop	aanbieden*krijg	aanbieden*kunne
aanbieden*leeftijd	aanbieden*leerwerk plek	aanbieden*licentie	aanbieden*lodier	aanbieden*lunch
aanbieden*maandag	aanbieden*maar	aanbieden*makelaars dienst	aanbieden*marines chip	aanbieden*mark
aanbieden*meer	aanbieden*menukaa rt	aanbieden*moet	aanbieden*muziekb estand	aanbieden*network
aanbieden*noodhulp	aanbieden*onderdak	aanbieden*ontslag	aanbieden*open	aanbieden*oplossen
aanbieden*pakket	aanbieden*petitie	aanbieden*piket	aanbieden*pil	aanbieden*pop
aanbieden*presenteerb lad	aanbieden*regeren	aanbieden*rei	aanbieden*serie	aanbieden*sloopbedri jf
aanbieden*soc	aanbieden*spel	aanbieden*spul	aanbieden*staag	aanbieden*stuk
aanbieden*tarief	aanbieden*tegen	aanbieden*telefonie	aanbieden*telefoon	aanbieden*televisie
aanbieden*tijd	aanbieden*veel	aanbieden*vergoede n	aanbieden*vermog ensdienst	aanbieden*verontsch uldigen
aanbieden*vervoer	aanbieden*verzekere n	aanbieden*virusbest rijder	aanbieden*vliegreis	aanbieden*vluchtverz ekeer
aanbieden*voeden	aanbieden*vol	aanbieden*voorstel	aanbieden*waarbij	aanbieden*was
aanbieden*wegens	aanbieden*werk	aanbieden*wil	aanbieden*worden	aanbieden*zaakwaar nemer
aanbieden*zaterdag	aanbieden*zijn	aanbieden*zonde	aanbieden*zullen	aanbieder*adsl
aanbieder*al	aanbieder*ander	aanbieder*behoren	aanbieder*buitenla nd	aanbieder*commerci eel
aanbieder*consumeren	aanbieder*divers	aanbieder*enkel	aanbieder*goedkoo p	aanbieder*gooi
aanbieder*groot	aanbieder*hebben	aanbieder*internettel efonie	aanbieder*kunne	aanbieder*make
aanbieder*marktaande el	aanbieder*meer	aanbieder*mobiel	aanbieder*nemen	aanbieder*nieuw
aanbieder*porno	aanbieder*prepaid	aanbieder*radio	aanbieder*steeds	aanbieder*succesvol
aanbieder*switch	aanbieder*telecom	aanbieder*toonaange vend	aanbieder*trendy	aanbieder*twee
aanbieder*veel	aanbieder*vier	aanbieder*voip	aanbieder*voornaa m	aanbieder*wereldwij d
aanbieder*worden	aanbieder*zijn	aanbieder*zoals	aanbieding*aantrek ken	aanbieding*af
aanbieding*afslaan	aanbieding*binnen	aanbieding*do	aanbieding*dozijn	aanbieding*hebben
aanbieding*introductie	aanbieding*komen	aanbieding*krijg	aanbieding*let	aanbieding*leuk
aanbieding*nieuw	aanbieding*persoonl ijk	aanbieding*play	aanbieding*rimpel	aanbieding*slaan
aanbieding*speciaal	aanbieding*tweejari g	aanbieding*verbetere n	aanbieding*voordel ig	aanbieding*waarmee

4. Conclusions

We have done a very basic task-based evaluation that has the advantage that it can be easily productized and does not put a heavy burden on processing text in a real environment. We have integrated the word combinations in Cornetto in a classification system that was built to assign IPTC topics to news documents. We found only a small improvement for the Cornetto-based index compared to the current system and system based on bigrams: the F-measure is 2.8% higher, recall is 4.5% higher and precision is 0.5% higher than the baseline. At the same time the indexes were only 1.2 times bigger by indexing word combinations using Cornetto. This means that it presents a realistic technology enhancement.

We expect that further improvements can be achieved through a more precise expansion. However, this requires development of further complex technologies and research to enable precise semantic mappings.