An Entity-Centric Coreference Resolution System for Person Entities with Rich Linguistic Information

Marcos Garcia and **Pablo Gamallo**
Centro Singular de Investigación en Tecnoloxías da Información (CiTIUS)
University of Santiago de Compostela
{marcos.garcia.gonzalez, pablo.gamallo}@usc.es

Abstract

This paper presents a first version of LinkPeople, an entity-centric system for coreference resolution of person entities. The approach combines (i) a multi-pass architecture which takes advantage of entity features at document-level with (ii) a set of linguistically-motivated constraints and rules which allows the system to restrict the candidates of a given mention. The paper includes evaluations and error analysis of LinkPeople in 3 different languages, achieving promising results (more than 81% F1 in different metrics). Both the system and the corpora are freely distributed.

1 Introduction

Coreference Resolution (CR) is a crucial task for several Natural Language Processing (NLP) applications such as Text Summarization, Machine Translation or Information Extraction (IE).

Specially for IE, person entities are those which more effort have deserved from different perspectives. Evaluations such as the Knowledge Base Population (KBP) Slot Filling Task (in the Text Analysis Conference)[1] and the Person Attribute Extraction (in the Web People Search Evaluation Campaign, WePS)[2], tasks such as Personal Name Matching (Cohen et al., 2003), or different works on Relation Extraction of person entities (Mann, 2002; Garcia and Gamallo, 2013) are some examples of their importance.

Recently, entity-centric models for coreference resolution, which use features from all the mentions of an entity, have shown better performance than pair-mention systems, which carry out coreference resolution on single pairs of mentions (Lee et al., 2013).[3] Furthermore, the use of linguistic information such as syntax or semantic knowledge has proved to be essential for high-precision CR (Ng and Cardie, 2002; Ponzetto and Strube, 2006; Uryupina, 2007).

This paper presents the first version of LinkPeople, an open-source system for CR of person entities. LinkPeople is inspired by the Stanford Deterministic Coreference Resolution System (Raghunathan et al., 2010; Lee et al., 2013), using a multi-pass architecture which applies a battery of modules sorted from high-precision to high-recall.

Moreover, the system presented in this paper adds new sieves based on linguistic knowledge, for both cataphoric and anaphoric mentions: It includes a high-precision module which finds cataphoric mentions of Noun Phrases (NP) and personal and elliptical pronouns. The inclusion of this module is based on the claim that definite NPs are not primarily anaphoric (Vieira and Poesio, 2000). In addition, LinkPeople applies a set of syntactic constraints on the pronominal CR module, increasing its precision by blocking links which do not satisfy the constraints (Mitkov, 1998; Palomar et al., 2001; Chaves and Rino, 2007).

The system was evaluated in three languages (Portuguese, Spanish and Galician) with promising results (F1 $\approx$ 83%, with BLANC score). Both LinkPeople and the corpora are freely distributed.[4]

[1] http://www.nist.gov/tac/data/index.html

[2] http://nlp.uned.es/weps/weps-3

[3] In this paper, a *mention* is every instance of reference to a person. An *entity* is the group of all the mentions referring to the same person in the text (Recasens and Martí, 2010).

[4] http://gramatica.usc.es/~marcos/coling14.tar.bz2

Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 741–752, Dublin, Ireland, August 23-29 2014.

Apart from this Introduction, Section 2 contains some related work. The architecture of the system is presented in Section 3 while its evaluation is shown in Section 4. Finally, the results of an error analysis are presented in Section 5, and some conclusions and further work are pointed out in Section 6.

2 Related Work

Coreference (and anaphora) resolution is one of the older topics in NLP, so it has been the subject of many works. Two main distinctions can be stated in coreference resolution systems: (i) mention-pair *vs* entity-centric approaches and (ii) machine learning-based *vs* rule-based models.

On the one hand, mention-pair systems classify two mentions in a text as coreferent or not, by using a feature vector obtained from this pair of mentions. On the other hand, entity-centric approaches determine if a mention (or a partial entity) belongs to another partial entity, using features from other mentions of the same (partial) entities.[5]

Machine learning classifiers for CR often use annotated corpora for training supervised models. Supervised models rely on these data in order to learn preferences and constraints (McCarthy and Lehnert, 1995; Soon et al., 2001; Ng and Cardie, 2002; Sapena et al., 2013), while unsupervised models apply clustering approaches to the coreference resolution problem (Haghighi and Klein, 2007; Ng, 2008).

Rule-based strategies make use of sets of rules and heuristics for finding the best element to link each mention to (Lappin and Leass, 1994; Baldwin, 1997; Mitkov, 1998; Bontcheva et al., 2002; Raghunathan et al., 2010; Lee et al., 2013). This last system is based on a multi-pass approach which first solves the *easy* links, then increasing the recall with more rules. Stoyanov and Eisner (2012) presented *EasyFirst*, which uses annotated corpora in order to know whether coreference links are easy or hard.

Concerning the languages LinkPeople deals with, some studies addressed pronominal CR in Portuguese (Paraboni, 1997; Chaves and Rino, 2007; Cuevas and Paraboni, 2008). Coelho and Carvalho (2005) adapted the Lappin and Leass (1994) algorithm for this language, while de Souza et al. (2008) presented a supervised approach for solving the coreference between NPs.

For Spanish, Palomar et al. (2001) presented a set of constraints and preferences for pronominal anaphora resolution. Recasens and Hovy (2009) analyzed the impact of several features for CR, then implemented in Recasens and Hovy (2010). The availability of a large coreference annotated corpus for Spanish (Recasens and Martí, 2010) also allowed other supervised systems being adapted for this language (Recasens et al., 2010).

To the best of our knowledge, there are no specific systems for coreference or anaphora resolution for Galician language.

Other related areas such as the above mentioned personal name matching perform coreference resolution of personal names by linking variants referring to the same person (Cohen et al., 2003).

The system presented in this paper uses a similar approach than Lee et al. (2013), adapting —and adding— some modules for person entities, and enriching others with linguistic-based heuristics such as cataphoric analysis and syntactic constraints.

3 Architecture of LinkPeople

LinkPeople is based on two main principles: (i) an entity-centric approach and (ii) a multi-pass architecture. On the one hand, the entity-centric approach allows the system to use all the features of an entity when a mention is evaluated. On the other hand, the multi-pass model dynamically enriches an entity (with new features) in every iteration. Thus, latter passes take advantage of the information provided by the previous coreference resolution modules.

Figure 1 shows a text with coreference annotation of person entities. It will be used to show how the system works. The input of LinkPeople needs to be pre-processed by NLP tools which provide PoS-tags, Named Entity Recognition (NER) and dependency analysis. In our experiments, FreeLing (Padró and Stanilovsky, 2012; Garcia and Gamallo, 2010) was used for tokenizing, lemmatizing and PoS-tagging. NER labeling for Spanish and Portuguese was also added by FreeLing (Carreras et al., 2003; Gamallo

[5]Partial entities are sets of mentions of the same entity.

Who was $_{1}$[the singer of the Beatles]$_{1}$. $_{2}$[The musician John Winston Ono Lennon]$_{1}$ was one of the founders of the Beatles. With $_{3}$[Paul McCartney]$_{2}$, $_{4}$[he]$_{1}$ formed a songwriting partnership. $_{5}$[Lennon]$_{1}$ was born at Liverpool Hospital to $_{6}$[Julia]$_{3}$ and $_{7}$[Alfred Lennon]$_{4}$. $_{8/9}$[$_{10}$[His]$_{1}$ parents]$_{3/4}$ named $_{11}$[him]$_{1}$ $_{12}$[John Winston Lennon]$_{1}$. $_{13}$[Lennon]$_{1}$ revealed a rebellious nature and acerbic wit. $_{14}$[The musician]$_{1}$ was murdered in 1980.

Figure 1: Example of a text with coreference annotation of person entities. Mentions appear inside brackets. Numbers on the left are mention ids, while entity ids appear in the right side.

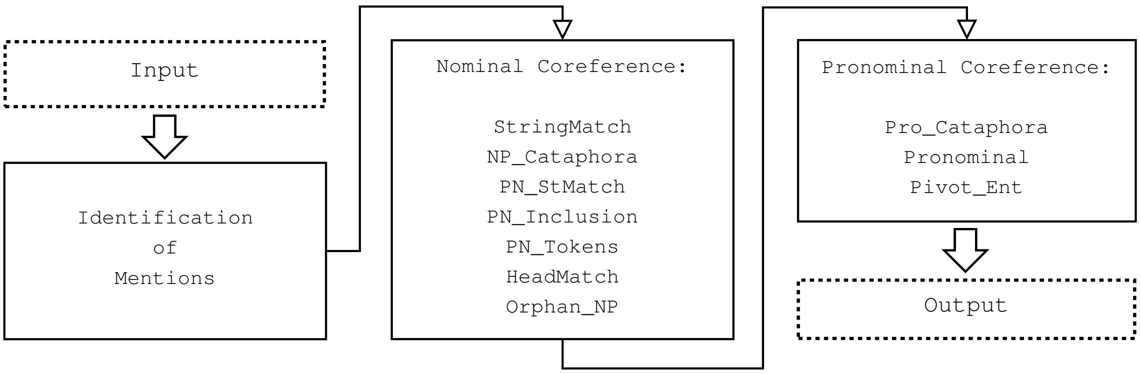

Figure 2: Architecture of the system.

and Garcia, 2011), while the named entities in Galician were classified by the system presented in Garcia et al. (2012). Finally, dependency information for the three languages was added by DepPattern (Gamallo and González López, 2011).

3.1 Coreference Resolution Modules

Figure 2 summarizes the architecture of the system, which starts by identifying the mentions. Then, a battery of nominal and pronominal CR modules is applied. Modules with high-precision are applied first, while other modules increase recall by taking advantage of the previously extracted features.

In the first stage, a specific pass identifies the mentions referring to a person entity, using the information provided by the PoS-tagger and the NER as well as applying basic approaches for NP and elliptical pronoun identification: First, personal names (and noun phrases including personal names) are identified. Then, it seeks for definite NPs whose head may refer to a person (e.g., “the singer”). Finally, this module selects singular possessives and applies basic rules for identifying relative, personal and elliptical pronouns (in sentence-initial position, after adverbial phrases and after preposition phrases) (Ferrández and Peral, 2000). At this step, each mention belongs to a different entity. Each entity contains the gender, number, head of a noun phrase, head of a Proper Noun (PN) and full proper noun as features. Once the mentions are identified, the coreference resolution modules are sequentially executed.

In order to perform CR, each module applies the following strategy (except for some exceptional rules, explained below): mentions are traversed from the beginning of the text and each one is *selected* if (i) it is not the first mention of the text and (ii) it is the first mention of its entity. Once a mention is selected, it looks backwards for *candidates* in order to find an appropriate antecedent (in the experiments, using the whole text). If an antecedent is found, mentions are merged together in the same entity. Then, the next selected mention is evaluated.

Besides the identification of mentions, current version of LinkPeople contains the following modules:

StringMatch (StM): this pass performs strict matching of the whole string of both mentions (the selected one and the candidate). In the example (Figure 1), mentions 13 and 5 are linked in this step.

NP_Cataphora (NP_C): this module verifies if the first mention —in the first paragraph— is an NP without a personal name. If so, it is considered a cataphoric mention, and the system checks if the next sentence contains a personal name as a Subject. In this case, these mentions are linked if they agree in gender and number. Mentions 1 and 2 in the example meet these requirements, so they merge. Note that, at the end of this pass, this entity has as NP heads the words ‘singer’ and ‘musician’, and ‘John Winston Ono Lennon’ as the PN. This module also matches fixed synonym structures through dependency paths, such as “Person$_A$, also known as Person$_B$”.

PN_StMatch (PN_St): in this stage, the system looks for mentions which share the whole PN, even if their heads are different (or if one of them does not have head). “The musician John Lennon” and “John Lennon” (not in Figure 1) would be an example.

PN_Inclusion (PN_I): here, the system verifies if the full PN of the selected mention (in the entity) includes the proper noun of the candidate mention (also in the entity), or vice-versa. In the example, mention 5 is linked to mention 2 in this step. Note that mention 7 is not linked to mention 5, because the full PN of mention 5 is now “John Winston Ono Lennon”, not compatible with “Alfred Lennon”. Also, mention 13 is not selected here because it is not the first mention of its entity.

PN_Tokens (PN_T): this module splits the full PN of a partial entity in their tokens, and verifies if the full PN of the candidate contains all the tokens in the same order, or vice-versa (except for some *stop words*, such as “Sr.”, “Jr.”, etc.). As the pair “John Winston Ono Lennon” - “John Winston Lennon” are compatible, mentions 12 and 5 are merged.

HeadMatch (HM): in this step, the system checks if the selected mention and the candidate one share the heads (or the heads of their entities). In Figure 1, mention 14 is linked to mention 13.

Orphan_NP (Orph): the last module of nominal CR applies a pronominal-based rule to orphan noun phrases. Here, a definite NP is marked as orphan if it is still a singleton and it does not contain a personal name. Thus, an orphan NP is linked to the previous PN with gender and number agreement. In the example, the mentions 8/9 are linked to 7 and 6.

Pro_Cataphora (Pro_C): similar to NP_Cataphora, this module verifies if a text starts with a personal (or elliptical) pronoun. If so, it looks in the following sentence if there are a compatible PN.

Pronominal (PRO): this is the standard module for pronominal CR. For each selected pronoun, it verifies if the candidate nominal mentions satisfy the syntactic (and morpho-syntactic) constraints (inspired by Palomar et al. (2001)). They include a set of constraints for each type of pronoun, which remove a candidate if any of the constraints is violated. Some of them are: an object pronoun (direct or indirect) cannot corefer with its subject (mention 11 *vs* mentions 8/9); a personal pronoun does not corefer with a mention inside a prepositional phrase (mention 4 *vs* mention 3), a possessive cannot corefer with the NP it belongs to (mention 10 *vs* mentions 8/9) or a pronoun prefers a subject NP as its antecedent (mentions 10 and 11 *vs* mentions 6 and 7). This way, in Figure 1 the pronominal mention 4 is linked to mention 2, and mentions 10 and 11 to mention 5. This module only looks in the same and previous sentence for candidates.

Pivot_Ent: this last module is only applied if there are orphan pronouns (not linked to any proper noun/noun phrase) at this step. First, it verifies if the text has a pivot entity, which is the more frequent personal name in a text whose frequency is at least 33% higher than the second person with more occurrences. Then, if there is a pivot entity, all the orphan pronouns are linked to its mention. If not, each orphan pronoun is linked to the previous PN/NP (with no constraint).

4 Evaluation

LinkPeople was tested on three different corpora (for Portuguese, Galician, and Spanish) with coreference annotation of person entities (Garcia and Gamallo, 2014). The annotation follows the SemEval-2010 guidelines. The corpus for Portuguese has about 51k tokens and $\approx$ 4,000 mentions. The Galician one, 42k tokens and $\approx$ 3,500 mentions. The Spanish corpus has over 46k tokens, and $\approx$ 4,500 mentions.

Some of the annotation (gender, number and syntactic labeling) was not manually revised, so it may contain errors (*regular setting*). The tests were carried out using a *gold mention* evaluation (i. e., using as input the corpora with the mentions already identified). Moreover, no external resources (gender dictionaries of proper nouns, WordNet, etc.) were used (*closed setting*).

In order to compare the results of LinkPeople, four well-known baselines were also evaluated: (i) *Singletons* (Stons), where every mention belongs to a different entity. (ii) *All_in_One* (AOne), where all the mentions belong to the same entity; (iii) *HeadMatch* (HMb), which clusters in the same entity mentions sharing the head and classify each pronoun as a singleton, and (iv) *HeadMatch_Pro* (HMP), same as the previous one, but linking each pronoun to the previous nominal mention with gender and number agreement.[6]

Five different metrics were taken into account: MUC (Vilain et al., 1995), B^3 (Bagga and Baldwin, 1998), $CEAF_{entity}$ (Luo, 2005), BLANC (Recasens and Hovy, 2011) and ConLL (Pradhan et al., 2011). They were computed with the scorers used in SemEval-2010[7] (for BLANC) and ConLL 2011[8] (for the other metrics).

Table 1 contains the results of the four baselines and of LinkPeople in the three corpora. The first block of each language includes the results of the baseline models. The central rows show the results of the different modules of LinkPeople (see Figure 2), added incrementally. The first nine rows (StM > PRO) include two default rules in order to classify mentions not covered by the active modules: (i) nominal mentions not analyzed are singletons and (ii) pronouns are linked to the previous nominal mention with gender agreement (except for those pronouns covered by PRO in this model). Furthermore, PRO systems do not restrict the number of previous sentences while looking for antecedents.

The last model (LinkP, the result of all the modules included in LinkPeople) does include a distance restriction in the Pronominal pass (see Section 3.1), so it combines Pronominal with Pivot_Ent modules.

As expected, *Singletons* and *HeadMatch* baselines produce poor results in most languages and metrics (*Singletons* values in MUC are null because this metric do not reward correctly identified singletons). However, *All_In_One* models achieved reasonable results in some scenarios (MUC and B^3). The differences between these values and those from SemEval-2010 are due to the existence (in this work) of just one type of entity. Journalistic and encyclopedic texts are often focused on just one or two persons, (i.e., there is a much lower number of entities in each text), so the precision is higher in *All_In_One* and lower in *Singletons*.

As Recasens and Hovy (2010) shown, *HeadMatch_Pro* baselines obtain good results in the three languages and with every metric ($\approx$ 60% and 67% in F1 BLANC and CoNLL, respectively).

Concerning the different passes of LinkPeople, the performance of the first matching modules depends on the distribution and structure of PNs and NPs in the corpora. In this respect, PN_StMatch works well in all the contexts. However, PN_Inclusion stands out in the Nominal modules, increasing in more than 5% (BLANC and CoNLL) the performance of the previous model. This is due to the high increase in recall together with the high-precision of this module.

It is worth noting that the addition of some modules seems to improve not only recall, but also precision. This is due to the execution of the two default rules: as the system uses more modules, the amount of (partial)entity mergings (usually) grows. Thus, the precision increases because the new mergings restrict incorrect links performed by the two default rules in the previous models.

HeadMatch module is the first one that deals with mentions without PN (except for the rules applied in NP_Cataphora, with low recall). Due to the knowledge provided by previous modules, it also benefits all the models and languages.

The performance of Orphan_NP and Pro_Cataphora also depends on the corpora and on the evaluation metric. The latter involves a 0.2% loss in Spanish with the BLANC score (but increases in 1.1% using CoNLL). However, Orphan_NP allows the system to not classify as singletons some mentions, which in turn helps to increase the performance of Pronominal modules. Similarly, Pro_Cataphora prevents the next sieve from selecting pronominal mentions that are cataphoric.

[6]Due to language differences and format issues, other coreference resolution systems could not be used for comparison (Raghunathan et al., 2010; Sapena et al., 2013).

[7]http://www.lsi.upc.edu/~esapena/downloads/scorer-v1.04.zip

[8]http://conll.cemantix.org/download/reference-coreference-scorers.v7.tar.gz

Lang	Model	MUC			B^3			$CEAF_e$			BLANC			CoNLL
		R	P	F1	R	P	F1	R	P	F1	R	P	F1	F1
Port.	Stons	-	-	-	15.0	100	26.1	65.3	10.9	18.7	50.0	29.0	36.7	14.9
	AOne	93.8	85.5	**89.4**	94.8	47.5	63.3	11.9	78.1	20.7	50.0	21.0	29.1	57.8
	HMb	26.5	93.9	41.3	22.2	97.9	36.2	72.3	16.1	26.4	53.6	78.5	44.2	34.6
	HMP	76.0	91.2	82.9	46.0	85.8	59.9	76.7	49.2	59.9	68.5	80.0	68.1	67.6
	StM	69.8	91.5	79.2	38.8	88.7	54.0	78.1	40.5	53.3	64.7	79.2	62.9	62.2
	NP_C	70.4	91.4	79.6	39.2	88.5	54.3	78.3	41.5	54.3	64.7	79.2	62.9	62.7
	PN_St	72.8	91.9	81.3	40.9	88.3	55.9	79.3	44.7	57.2	65.0	79.2	63.4	64.8
	PN_I	77.1	92.5	84.1	50.5	87.5	64.0	81.9	52.7	64.1	71.1	81.0	71.2	70.8
	PN_T	77.3	92.5	84.2	50.8	87.5	64.3	82.0	53.0	64.4	71.1	81.0	71.3	71.0
	HM	79.7	92.3	85.6	53.6	85.5	65.9	81.3	58.3	67.9	71.5	80.7	71.7	73.1
	Orph	83.4	91.8	87.4	58.1	82.7	68.3	81.4	70.2	75.4	71.6	80.3	71.9	77.0
	ProC	83.4	91.8	87.4	58.1	82.7	68.3	81.4	70.3	**75.5**	71.6	80.3	72.0	77.0
	PRO	81.8	91.7	86.4	59.1	83.9	69.3	82.7	66.5	73.7	76.0	83.7	76.7	76.5
	LinkP	82.7	92,7	87.4	65.8	84.5	**74.0**	84.4	67.9	75.2	83.6	85.4	**84.2**	**78.9**
Gal.	Stons	-	-	-	14.6	100	25.4	71.7	11.0	19.1	50.0	28.4	36.3	14.8
	AOne	96.6	86.0	91.0	97.1	53.9	69.3	9.0	82.7	16.2	50.0	21.6	30.1	58.8
	HMb	21.1	90.5	34.2	20.2	97.5	33.5	74.1	14.3	24.0	51.3	74.7	39.1	30.6
	HMP	81.9	89.8	85.7	44.1	83.6	57.7	70.0	53.5	60.6	61.3	76.5	57.9	68.0
	StM	77.1	90.6	83.3	36.5	86.7	51.4	75.1	45.5	56.6	58.9	76.9	53.7	63.8
	NP_C	77.6	90.7	83.6	37.2	86.7	52.1	75.2	46.2	57.3	59.2	77.0	54.3	64.3
	PN_St	79.0	90.9	84.6	39.1	86.2	53.8	75.6	48.8	59.3	59.7	77.0	55.1	65.9
	PN_I	83.1	91.5	87.1	46.7	85.3	60.4	76.7	57.8	66.0	62.5	77.5	59.5	71.1
	PN_T	83.3	91.5	87.2	48.2	85.3	61.6	76.9	58.6	66.5	63.2	77.9	60.5	71.8
	HM	84.6	91.6	87.9	49.8	84.4	62.6	76.8	62.0	68.6	63.4	77.5	60.8	73.1
	Orph	84.7	91.3	87.9	49.9	83.9	62.6	76.8	63.2	69.4	63.3	77.3	60.8	73.3
	ProC	84.7	91.3	87.9	49.1	83.9	62.6	76.8	63.2	69.4	63.3	77.3	60.8	73.3
	PRO	86.9	92.5	89.6	60.7	86.8	71.4	82.8	72.2	77.1	73.6	82.0	73.9	79.4
	LinkP	89.0	94.6	**91.7**	72.9	88.4	**79.9**	87.6	76.6	**81.7**	82.7	85.8	**83.4**	**84.4**
Spa.	Stons	-	-	-	10.9	100	19.7	69.5	8.7	15.4	50.0	29.4	37.0	11.7
	AOne	91.7	88.4	**90.0**	92.6	51.3	66.0	6.4	83.0	11.9	50.0	20.6	29.2	55.9
	HMb	20.7	94.2	34.0	15.4	98.0	26.6	75.4	11.9	20.6	51.3	74.6	39.9	27.0
	HMP	78.2	90.7	84.0	35.3	81.2	49.2	72.9	51.5	60.4	59.3	74.7	55.5	64.5
	StM	73.9	90.7	81.4	30.1	83.7	44.3	73.9	41.6	53.3	58.6	75.6	54.1	59.7
	NP_C	74.1	90.7	81.5	30.2	83.7	44.4	73.9	42.0	53.6	58.6	75.6	54.1	59.8
	PN_St	75.4	91.0	82.5	31.2	83.1	45.4	73.8	44.1	55.2	58.6	75.4	54.3	61.0
	PN_I	78.8	91.7	84.8	39.3	82.2	53.1	75.9	52.8	62.3	62.0	76.7	59.6	66.7
	PN_T	79.0	91.7	84.9	40.0	82.1	53.8	76.0	53.3	62.7	62.6	76.3	60.5	67.1
	HM	80.5	92.0	85.9	41.7	80.9	55.1	75.6	57.3	65.2	63.1	75.0	61.4	68.7
	Orph	81.1	91.9	86.1	42.3	80.5	55.5	75.4	59.8	66.7	63.2	75.0	61.6	69.4
	ProC	82.3	91.9	86.8	43.2	79.6	56.0	74.6	64.1	68.9	63.0	74.7	61.4	70.6
	PRO	82.6	92.4	87.2	46.0	80.8	58.7	77.5	65.8	71.2	66.8	77.9	66.2	72.4
	LinkP	84.1	94.1	88.8	62.9	84.8	**72.2**	83.4	71.0	**76.7**	81.7	84.9	**82.6**	**79.2**

Table 1: Results of LinkPeople compared to the baselines in Portuguese (Port.), Galician (Gal.) and Spanish (Spa.). *LinkP* contains the results of the execution of the whole system.

The standard pronominal resolution module also increases the accuracy of all the systems (with the only exception in Portuguese language with the CoNLL score, which also had a high increase with the Orphan_NP module).

Finally, one of the main contributions to the performance of LinkPeople is the combination of the Pronominal module with the Pivot_Ent one. This combination reduces the scope of the Pronominal module, thus strengthening the impact of syntactic constraints. Furthermore, Pivot_Ent looks for a prominent person entity in each text, and links the orphan pronouns to this entity. In the three languages, the improvement is noticeably better with the BLANC score.

Last row of each language shows the current results of LinkPeople in the three corpora, with macro-average values of $\approx 83\%$ and $\approx 81\%$ with BLANC and CoNLL scores, respectively.

5 Error Analysis

In order to determine the major classes of errors produced by the system, 150 errors (50 for each language) were randomly selected from the output of LinkPeople. Each error was analyzed in order to find its source, and was classified according to its typology. This section shows the different error typologies together with some examples, sorted by their frequency in the corpora (first percentage in parenthesis is the average frequency, while the other three correspond to Portuguese, Galician and Spanish values, respectively).[9] They are real examples of incorrectly analyzed mentions (or pairs of mentions belonging to the same entity), with some simplifications due to space reasons:

5.1 Missing links between Noun Phrases and/or Proper Nouns (46%: 58% / 32% / 48%)

This category includes some error typologies that differ in the type of knowledge and analysis required by the system in order to accurately link two mentions:

Synonym heads (35.3%: 48% / 32% / 26%): The most frequent type of missing links was produced by mentions of the same entity whose heads are synonyms:

Mention A: "El *joven*" (the *young*)
Mention B: "el *muchacho*" (the *boy*)

External (real-world) knowledge (6%: 0% / 0% / 18%): This class includes mentions of the same entity which do not share the lexical features, usually because they refer to well-known entities in the real world:

Mention A: "la presidenta" (the president)
Mention B: "Cristina Kirchner"

Here, the noun phrase "the president" is used to refer "Cristina Kirchner", but the mentions are not linked because the system does not take advantage of resources that define Cristina Kirchner as a *president*.

Semantic knowledge (2.7%: 4% / 0 % / 4%): Lack of other type of semantic knowledge, such as hyponym-hypernym pairs, also involves missing links like the following:

Mention A: "o escocês" (the scotish)
Mention B: "o britânico" (the british)

Head modifiers (1.3%: 4% / 0 % / 0%): Internal modifiers of some heads may also produce missing links, as in the following example, where a mention does not contain the modifier *adjunto* (vice):

Mention A: "o *ministro* (the *minister*)
Mention B: "o *ministro-adjunto*" (the *vice-minister*)

Spelling differences (0.7%: 2% / 0% / 0%): Some personal names are spelled differently in the same text:

Mention A: "André *Villas-Boas*"
Mention B: "André *Villas Boas*"

5.2 Errors due to incorrect predicted (syntactic and morpho-syntactic) analysis (15.3%: 2% / 22% / 22%)

Since the corpora do not have PoS-tagging and dependency labels fully revised, some of these errors involve missing and spurious links between mentions.

Errors in syntactic constraints (10.7%: 0% / 16% / 16%): Direct and indirect object pronouns incorrectly labeled are not covered by some of the syntactic constraints, thus involving an incorrect link between a pronoun and its subject noun phrase.

[9] The results of 0% in some languages and categories do not mean that these languages cannot have those error typologies, but they did not appear due to the small number of errors evaluated.

Incorrect gender (2.7%: 2% / 4% / 2%): The gender of some nouns and adjectives also can be wrongly labeled, so other mentions may be incorrectly linked, or involve a missing link. For instance, the word *atleta* (sportsperson, which can be both masculine or feminine), labeled as masculine blocked a link to the feminine pronoun *ela* (she) in Galician.

Incorrect head (2%: 0% / 2% / 4%): Errors in PoS-tagging (usually between nouns and adjectives) also produce wrong dependency analysis, which in turn involve incorrect extractions of the NP heads:

Mention: "el jugador alemán" (the german player)
Extracted Head: *alemán (*german, instead of *jugador/player*)

5.3 Missing links due to long distance pronominal anaphora (11.3%: 14% / 18% / 2%)

This kind of errors arises when the distance between a pronoun and its nominal antecedent is outside the scope of a rule (in our case, between two and four sentences, depending on the module), and the antecedent is not the pivot entity.

5.4 Errors due to quoted speech coreference (10%: 10% / 14% / 6%)

Another category of errors includes mentions inside quoted speech. These mentions can refer to the speaker (first person) or to a third person in the quoted speech:

First person (4.7%: 6% / 6% / 2%): The 1st person of the quoted speech should be linked to the speaker instead of to a previous entity (note that the elliptical pronoun might also be a 3rd person pronoun):

"Si $\emptyset_{1^{st}}$ tuviera que redactar [...]", resumió Lezcano$_{\text{Speaker}}$.
"If [I$_{1^{st}}$] had to write [...]", Lezcano$_{\text{Speaker}}$ summarized.

Third person (5.3%: 4% / 8% / 4%): 3rd persons of a quoted speech should not be linked to the speaker:

Gustavo$_{\text{Speaker}}$: "Cuando yo$_{1^{st}}$ me fui, él$_{3^{rd}}$ dejó Boca."
Gustavo$_{\text{Speaker}}$: "When I$_{1^{st}}$ quit, he$_{3^{rd}}$ left Boca."

5.5 Spurious links in plural mentions (5.3%: 4% / 4% / 8%)

Coreference of plural mentions was performed through basic links to the previous entities, producing incorrect classifications. Also, some plural mentions include entities with different genders (e.g., *amigos* —friends— may refer to feminine and masculine entities, but the grammatical gender of the word is masculine in the three analyzed languages):

$_1$[Hulk]$_1$, $_2$[Moutinho]$_2$ e $_3$[Álvaro Pereira]$_3$ na lista de compra de $_4$[Villas-Boas]$_4$ [...]. $_{5/6/7}$[O trio do F.C. Porto]$_{2/3/*4}$ [...].

$_1$[Hulk]$_1$, $_2$[Moutinho]$_2$ and $_3$[Álvaro Pereira]$_3$ in the shopping list of $_4$[Villas-Boas]$_4$ [...]. $_{5/6/7}$[The F.C. Porto trio]$_{2/3/*4}$ [...].

In this example, the plural mention (*O trio do F.C. Porto*) is linked to the previous nominal mentions with gender agreement, so an incorrect link between mentions 7 and 4 is done.

5.6 Errors due to incorrect gender agreement (4.7%: 4% / 4% / 6%)

Some nominal phrases referring to the same entity may have different gender, thus causing wrong links:

Mention A: la víctima (the victim: feminine)
Mention B: el muchacho (the boy: masculine)

5.7 Errors produced by constraints and Pivot_Ent modules (4.6%: 6% / 0% / 8%)

The syntactic constraints, although precise, may restrict some correct links. This can involve (i) an incorrect discourse analysis or (ii) the application of Pivot_Ent, linking the mention to the most frequent entity, which might be incorrect:

$_1$[El escritor]$_1$ tuvo que visitar a $_2$[Martín]$_2$ en el hotel. Según $_3\emptyset_{*1}$ dijo [...]
$_1$[The writer]$_1$ had to visit $_2$[Martín]$_2$ in the hotel. As $_3$[he]$_{*1}$ said [...]

Here, the elliptical subject of *dijo* (said) is *Martín*, but the link is blocked due to a syntactic constraint: the antecedent of the (subject) elliptical pronoun should be a subject. Thus, the system incorrectly links mention 3 to mention 1.

5.8 Spurious links between Noun Phrases sharing the same head (1.3%: 0% / 4% / 0%)

In the same text, different entities can share their heads in some mentions, which may involve errors in coreference links, depending on their position and on their features. Thus, the NP "the president" may be linked to two different persons like "the president of the Academia" and "the president of the Government".

5.9 Spurious links produced by errors in previous modules (0.7%: 0% / 2% / 0%)

First modules also produce some incorrect clusters which involve errors in further modules. For instance, in the Galician corpus, NP_Cataphora incorrectly linked the noun phrase *o alcalde* (the mayor) to the proper noun "Dorribo". Then, HeadMatch merged "Dorribo" with *o alcalde Orozco*, creating an incorrect entity that contains two different persons (Dorribo and Orozco).

5.10 Errors due to fixed language structures (0.7%: 2% / 0% / 0%)

Other minor errors include some fixed structures such as the following cataphoric possessive:

Por $_1$[sua]$_1$ parte, $_2$[Cristina]$_{*2}$ [...]
For $_1$[her]$_1$ part, $_2$[Cristina]$_{*2}$ [...]

The results of the error analysis bring interesting information to further work. Thus, including some kind of semantic knowledge (synonyms), improving pronominal coreference resolution and implementing specific rules for quoted speech might solve many of the most frequent errors made by LinkPeople.

6 Conclusions and Further Work

This paper presents the first version of LinkPeople, an open-source entity-centric approach for coreference resolution of person entities which applies a battery of deterministic modules enriched with precise linguistic information.

The system was evaluated in three different languages (Portuguese, Galician and Spanish), clearly surpassing some powerful baselines and achieving promising results.

The addition of rules focused on cataphoric coreference as well as pronominal constraints based on syntactic and discourse restrictions increases the performance of similar approaches with lack of this kind of knowledge.

Current work explores better nominal (Elsner and Charniak, 2010) and pronominal constraints and dedicated handling of plural mentions. In further work, the implementation of an inheritance constraint is planned, which could prevent the merging of partial entities if their mentions were blocked by previous modules. Moreover, the extension of the system for solving the coreference of other types of entities is also planned.

Acknowledgments

This work has been supported by the HPCPLN project – Ref: EM13/041 (Galician Government) and by the Celtic – Ref: 2012-CE138 and Plastic – Ref: 2013-CE298 projects (Feder-Interconnecta).

References

Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *Proceedings of the Workshop on Linguistic Coreference at the International Language Resources and Evaluation Conference (LREC 1998)*, volume 1, pages 563–566.

Breck Baldwin. 1997. CogNIAC: high precision coreference with limited knowledge and linguistic resources. In *Proceedings of a Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*, pages 38–45. Association for Computational Linguistics.

Kalina Bontcheva, Marin Dimitrov, Diana Maynard, Valentin Tablan, and Hamish Cunningham. 2002. Shallow Methods for Named Entity Coreference Resolution. In *Proceedings of the Workshop on Chaınes de références et résolveurs d'anaphores at Traitement Automatique des Langues Naturelles (TALN 2002)*.

Xavier Carreras, Lluís Márquez, and Lluís Padró. 2003. A simple named entity extractor using AdaBoost. In *Proceedings of the 7th Conference on Computational Natural Language Learning (CoNLL 2003): Shared Task*, volume 4, pages 152–155. Association for Computational Linguistics.

Amanda Rocha Chaves and Lucia Helena Machado Rino. 2007. A resolução de pronomes anafóricos do português com base em heurísticas que apontam o antecedente. In *Proceedings of VI Congresso de Pós-Graduação da UFSCar*, volume 2, pages 1272–1273, São Carlos, São Paulo.

Thiago Thomes Coelho and Ariadne Maria Brito Rizzoni Carvalho. 2005. Uma adaptação do algoritmo de Lappin e Leass para resolução de anáforas em português. In *III Workshop em Tecnologia da Informação e da Linguagem Humana–TIL. Proceedings of XXV Congresso da SBC*, pages 2069–2078.

William W. Cohen, Pradeep Ravikumar, and Stepehn G. Fienberg. 2003. A Comparison of String Distance Metrics for Name-Matching Tasks. In *Proceedings of the IJCAI 2003 Workshop on Information Integration on the Web*, pages 73–78.

Ramon Ré Moya Cuevas and Invandré Paraboni. 2008. A machine learning approach to portuguese pronoun resolution. In *Advances in Artificial Intelligence (IBERAMIA 2008)*, pages 262–271. Springer-Verlag.

José Guilherme C. de Souza, Patrícia Gonçalves, and Renata Vieira. 2008. Learning Coreference Resolution for Portuguese Texts. In *Computational Processing of the Portuguese Language (PROPOR 2008)*, pages 153–162. Springer-Verlag.

Micha Elsner and Eugene Charniak. 2010. The same-head heuristic for coreference. In *Proceedings of the 48th Association for Computational Linguistics Conference Short Papers (ACL 2010)*, pages 33–37. Association for Computational Linguistics.

Antonio Ferrández and Jesús Peral. 2000. A computational approach to zero-pronouns in Spanish. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics (ACL 2000)*, pages 166–172. Association for Computational Linguistics.

Pablo Gamallo and Marcos Garcia. 2011. A resource-based method for named entity extraction and classification. In *Progress in Artificial Intelligence (LNCS/LNAI)*, volume 7026/2011, pages 610–623, Berlin. Springer-Verlag.

Pablo Gamallo and Isaac González López. 2011. A Grammatical Formalism Based on Patterns of Part-of-Speech Tags. *International Journal of Corpus Linguistics*, 16(1):45–71.

Marcos Garcia and Pablo Gamallo. 2010. Análise Morfossintáctica para Português Europeu e Galego: Problemas, Soluções e Avaliação. *Linguamática*, 2(2):59–67.

Marcos Garcia and Pablo Gamallo. 2013. Exploring the Effectiveness of Linguistic Knowledge for Biographical Relation Extraction. *Natural Language Engineering*. Available on CJO 2013 doi:10.1017/S1351324913000314.

Marcos Garcia and Pablo Gamallo. 2014. Multilingual corpora with coreferential annotation of person entities. In *Proceedings of the 9th edition of the Language Resources and Evaluation Conference (LREC 2014)*, pages 3229–3233. European Language and Resources Association.

Marcos Garcia, Iria Gayo, and Isaac González López. 2012. Identificação e Classificação de Entidades Mencionadas em Galego. *Estudos de Lingüística Galega*, 4:13–25.

Aria Haghighi and Dan Klein. 2007. Unsupervised coreference resolution in a nonparametric bayesian model. In *Proceedings of the 45th Annual Meeting on Association for Computational Linguistics (ACL 2007)*, volume 45, pages 848–855. Association for Computational Linguistics.

Shalom Lappin and Herbert J. Leass. 1994. An algorithm for pronominal anaphora resolution. *Computational linguistics*, 20(4):535–561.

Heeyoung Lee, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2013. Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics*, 39(4):885–916.

Xiaoqiang Luo. 2005. On Coreference Resolution Performance Metrics. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*, pages 25–32. Association for Computational Linguistics.

Gideon S. Mann. 2002. Fine-grained proper noun ontologies for question answering. In *Proceedings of the 2002 workshop on Building and using semantic networks*, volume 11. Association for Computational Linguistics.

Joseph McCarthy and Wendy G. Lehnert. 1995. Using decision trees for coreference resolution. In *Proceedings of the 14th International Conference on Artificial Intelligence*, pages 1050–1055.

Ruslan Mitkov. 1998. Robust pronoun resolution with limited knowledge. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (ACL/COLING 1998)*, volume 2, pages 869–875. Association for Computational Linguistics.

Vincent Ng and Claire Cardie. 2002. Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL 2002)*, pages 104–111. Association for Computational Linguistics.

Vincent Ng. 2008. Unsupervised models for coreference resolution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2008)*, pages 640–649. Association for Computational Linguistics.

Lluís Padró and Evgeny Stanilovsky. 2012. FreeLing 3.0: Towards Wider Multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Turkey. European Language and Resources Association.

Manuel Palomar, Antonio Ferrández, Lidia Moreno, Patricio Martínez-Barco, Jesús Peral, Maximiliano Saiz-Noeda, and Rafael Muñoz. 2001. An algorithm for anaphora resolution in Spanish texts. *Computational Linguistics*, 27(4):545–567.

Ivandré Paraboni. 1997. Uma arquitetura para a resolução de referências pronominais possessivas no processamento de textos em língua portuguesa. Master's thesis, Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre.

Simone Paolo Ponzetto and Michael Strube. 2006. Exploiting semantic role labeling, WordNet and Wikipedia for coreference resolution. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (HLT/NAACL 2006)*, pages 192–199. Association for Computational Linguistics.

Sameer Pradhan, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue. 2011. CoNLL-2011 Shared Task: Modeling Unrestricted Coreference in OntoNotes. In *Proceedings of the 15th Conference on Computational Natural Language Learning (CoNLL 2011): Shared Task*, pages 1–27. Association for Computational Linguistics.

Kathik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky, and Christopher Manning. 2010. A multi-pass sieve for coreference resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP 2010)*, pages 492–501. Association for Computational Linguistics.

Marta Recasens and Eduard Hovy. 2009. A deeper look into features for coreference resolution. In *Anaphora Processing and Applications*, pages 29–42. Springer-Verlag.

Marta Recasens and Eduard Hovy. 2010. Coreference resolution across corpora: Languages, coding schemes, and preprocessing information. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*, pages 1423–1432. Association for Computational Linguistics.

Marta Recasens and Eduard Hovy. 2011. BLANC: Implementing the Rand Index for Coreference Evaluation. *Natural Language Engineering*, 17(4):485–510.

Marta Recasens and M. Antònia Martí. 2010. AnCora-CO: Coreferentially annotated corpora for Spanish and Catalan. *Language Resources and Evaluation*, 44.4:315–345.

Marta Recasens, Lluís Màrquez, Emili Sapena, M. Antònia Martí, Mariona Taulé, Véronique Hoste, Massimo Poesio, and Yannick Versley. 2010. SemEval-2010 Task 1: Coreference resolution in multiple languages. In *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval'10)*, pages 1–8. Association for Computational Linguistics.

Emili Sapena, Lluís Padró, and Jordi Turmo. 2013. A Constraint-Based Hypergraph Partitioning Approach to Coreference Resolution. *Computational Linguistics*, 39(4).

Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational linguistics*, 27(4):521–544.

Veselin Stoyanov and Jason Eisner. 2012. Easy-first coreference resolution. In *Proceedings of the International Conference on Computational Linguistics (COLING 2012)*, pages 2519–2534.

Olga Uryupina. 2007. *Knowledge acquisition for coreference resolution*. Ph.D. thesis, Universität des Saarlandes.

Renata Vieira and Massimo Poesio. 2000. An empirically based system for processing definite descriptions. *Computational Linguistics*, 26(4):539–593.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Proceedings of Message Understanding Conference 6 (MUC-6)*, pages 45–52. Association for Computational Linguistics.