Hybrid Approach to Zero Subject Resolution for multilingual MT

- Spanish-to-Korean Cases -

Arum Park
Dept. of German Linguistics & Literature,
Sungkyunkwan University /
25-2, Sungkyunkwan-Ro, Jongno-Gu,
Seoul, Korea
remin2@skku.edu

Munpyo Hong*
Dept. of German Linguistics & Literature,
Sungkyunkwan University /
25-2, Sungkyunkwan-Ro, Jongno-Gu,
Seoul, Korea
skkhmp@skku.edu

Abstract

The current paper proposes a novel approach to Spanish zero pronoun resolution in the context of Spanish to Korean Machine Translation (MT). Spanish is one of the well-known 'pro-drop' languages so that especially a subject pronoun is often omitted, if it can be inferred from the linguistic as well as non-linguistic context. In Spanish to Korean MT the omitted subject doesn't need to be restored in many cases as Korean also allows a zero subject. However, there are some cases where the omitted subject must be identified to ensure a correct translation. To restore the omitted subject, linguistic clues can be employed, as Spanish verbs undergo morphological flections with respect to the gender and number. However, there still remain some ambiguous cases in which there are more than two possible subject candidates for the specific verb endings. In this paper, we propose a hybrid approach to resolve Spanish zero subject that integrates linguistic knowledge (morphological information) and artificial intelligence knowledge (machine learning approach). We proposed 11 linguistically motivated features for ML (Machine Learning). Our approach has been implemented with WEKA 3.6.10 and evaluated by using 10 fold cross validation method. The accuracy of the proposed method reached 83.6% while the baseline method that randomly chooses a possible subject candidate among three most frequent subject types shows only 33.3% accuracy rate.

1 Introduction

Spanish is one of the so-called pro-drop languages where certain pronouns may be omitted. In Spanish, the pronominal subject can be deleted and is called a zero subject. A zero subject is the most frequent type of anaphoric expression in Spanish.[1] Palomar et al.(2001) reported that about 65.5% of the pronouns are the zero subject pronoun in pronoun occurrences in Spanish corpus.

Spanish zero subject is one of the important issues that must be tackled in Spanish-to-Korean MT. This kind of pronoun is very important due to its high frequency in Spanish texts. In many cases its resolution is obligatory in Spanish-to-Korean MT.

Let us consider the following example. In this example, the omitted subject is represented by the symbol ø.

(1) Luis quiere que **Ø** vayamos[1st plural] a la playa.
Luis want that go to the beach

* Corresponding author
[1] Palomar et al.(2001)

28th Pacific Asia Conference on Language, Information and Computation pages 254–261

lwuisunun wulituli haypyeney kakilul wunhanta

"Luis wants us to go to the beach."

In Spanish, the subject pronoun and the verb must agree in person and number. Even though the subject pronoun is not present in the sentence, the zero subject can be restored from the verb ending, as '-os' is a 1st person plural morpheme. This gives us clues to resolve a Spanish zero subject.

However, there is another case for which to use verb ending for Spanish zero subject resolution is not enough. In Spanish, the verb ending for the 3rd person singular subject 'él(he)', 'ella(she)' and formal 2nd person singular subject 'Usted(you)' is same and so are for the plural subjects. Also for some verbs in a specific tense like *pretérito imperfecto*, the verb endings for the 1st and 3rd person singular are identical. Even in other sentence mood, there are some verbs which conjugate in the same way. Therefore, there still exists a problem to select the one right subject among possible subjects for some verb endings. For these cases, we need to suggest another method to select the one right subject among other possible subjects.

We introduce a machine learning method to resolve the case in which morphological information is not enough to resolve Spanish zero subject. Machine learning (ML) has already been successfully used in the computational linguistics for disambiguation and classification issues. Selecting one right subject among possible candidates can also be regarded as a disambiguation issue.

In this paper we propose a hybrid approach to zero subject in Spanish, which combines linguistic knowledge and ML approach in one model. The hybrid approach can benefit from the strengths of both approaches.

The related works about anaphora resolution and their limitation are presented in Section 2. In Section 3, we suggest the method for Spanish zero subject resolution. 11 features for ML method are also proposed. In Section 4, the effect of using ML is evaluated. Finally, the conclusion is presented in Section 5.

2 Related works

A zero pronoun subject has drawn much attention for various applications in the computational linguistics. Both rule-based and machine learning approaches have been utilized for languages such as Japanese (Okumura, M. and K. Tamura., 1996), Chinese (Zhao, S. and H.T. Ng., 2007), Korean (Han, N., 2004) and Spanish (Ferrández, A. and J. Peral., 2000) for zero pronoun identification and resolution.

The current anaphora resolution approaches rely mostly on linguistic knowledge in a rule-based framework. They try to find the right antecedent for the anaphora employing constraints and preference for the resolution.

The constraints discard some of the antecedent candidates for the anaphora and tend to be absolute. Morphological information is one of the constraints. For example, pronominal anaphors and antecedents must agree in person, gender, and number (e.g. Rich, E. and S. LuperFoy., 1988; Carbonell, J.G. and R.D. Brown., 1988).

Semantic information such as semantic consistency is also used as a constraint (e.g. Wilks, Y., 1973). This constraint stipulates that if satisfied by the anaphor, the semantic consistency constraint must also be satisfied by its antecedent. Although using constraints is the surest way to remove non-anaphoric pairs, they are not always sufficient to distinguish between a set of possible candidates.

Preference is a heuristic rule and it sets priorities in the list of the antecedent candidates which are left after constraints were applied to the list. Some of the works using preference are based on the centering theory. Centering theory (e.g. Brennan et al. 1987; Dahl, D.A. and C.N. Ball., 1990; Mitkov, R., 1994; Sidner, C., 1986; Stys, M.E. and S.S. Zemke., 1995; Walker et al. 1994) is a kind of preference rule because it gives more preference to certain candidates and less to others in forward-looking center lists.

However, there seems to be some difficulty in applying the centering theory for Spanish zero subject resolution. The text type we focused on is a spoken Spanish, so that there are even no antecedents for some zero subjects in the text. Therefore, to make a list of forward-looking centers would be difficult in Spanish spoken texts.

Rello et al.(2010) dealt with ML for Spanish anaphora phenomenon but did not focus on Spanish zero subject resolution. In zero anaphora resolution, non-referential subject ellipses need to be filtered out. They present a three-fold classification of subjects as (1) explicit and

referential (2) elliptic and referential (zero pronouns) and (3) elliptic and non-referential (impersonal constructions) using ML techniques.
Unlike this work, the aim of our work lies in resolving zero pronouns, not in classifying the subject types. The focus of our work is only on the subject class (2) in Rello et al.(2010).

3 Methodology

In Spanish, morphological information such as person and number agreement is a certain constraint to discard wrong antecedent candidates. According to the result of our experiment, in about 70% of the sentence, a zero subject can be restored by consulting the verbal flection information as can be seen in (2).

(2) Ø Estudias[2nd person sing] español?
study Spanish

nenun supheyinelul kongpuhani

"Do you study Spanish?"

The rest of the sentences, about 30% are the cases where using verb ending is not enough to restore one right subject. As for some verbs, multiple subjects can be possible candidates for the zero subject as in (3)-(5).

(3) ¿Ø Podría[3rd person sing] llegar tarde?
seem come late

(ku/kunye/tangsin) nuckey ol kes kathayo

"Does(Do) he/she/you seem to come late?

(4) ¿Ø Porqué iba[1st&3rd person sing] a ir?
why be going to go

(na/ku/kunye/tangsin) oway kalyeko haysseyo

"Why were(was) I/he/she/you going to go?"

(5) Ø Está[2nd&3rd person sing] en periodo de prueba.
be for a while probation

(ne/ku/kunye/tangsin) tangpwunkan kunsiniya

"You/He/She/You were(was) placed under probation for a while."

In these cases, there still remains a problem that one right subject among possible subjects has to be selected. For the cases, we applied ML method to select the right one for the zero subject.
Using only ML method without linguistic information is not the most optimal approach to resolve zero subjects. Applying the constraint is the surest way to narrow down the list of candidate subjects. For this reason, we proposed a 'hybrid approach' to Spanish zero subject resolution, which combines linguistic knowledge with ML. Our proposal is presented in Figure 1.

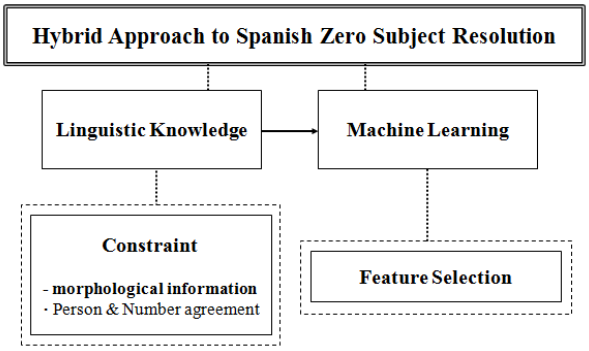

Figure 1. Hybrid approach to Spanish zero subject resolution

4 Experiments

In order to use a ML method, 11 features for Spanish zero subject resolution we introduced are presented in Table 1. The features were selected according to their linguistic as well as non-linguistic relevance to zero pronouns.

Feature Type	Feature		Value
Morpho-logical/ Syntactic	*f1*	Syntactic function of antecedent	(sub), (obj-v), (obj-p), (pos-adj), (ref-pro), (voc), (none)[2]
	f2	Person of the verb	the third person(1), the first/third person(2), the third person-indicativo/the second person-imperativo(3)
	f3	Number of the verb	singular(1), plural(2)

[2] 'none' represents the cases where there are no antecedent in the case of extra-sentential zero pronoun types.

	f4	Gender of antecedent	masculine(1), feminine(2), neutral(3)[3], (none)
Semantic	*f5*	Semantic class of antecedent	person (0), object (1), others(2), (none)
Relational	*f6*	Distance	in the same sentence(0), 1 sentence before(1) and 2 sentences before(2) and so on, (none)
Specific to Spanish	*f7*	Presence of a possessive adjective in the same sentence (same as coreferent)	false(0), possessive adjective in the first person (1), possessive adjective in the third person(2)
	f8	Presence of a reflexive pronoun in the same sentence (same as coreferent)	false(0), reflexive pronoun in the first person (1), reflexive pronoun in the third person (2)
	f9	Presence of antecedent	false(0), true(1)
	f10	mood of sentence	indicative(1), conditional(2), imperative(3), subjunctivo(4)
	f11	tense of sentence	PERFECTO(3), FUTURO IMPERFECTO(4), PRET.PERFECTO(5), PRET. PLUSCUAMPLERFECTO (6), FUTURO PERFECTO(7)

Table 1. 11 features for ML

There are features that are related to syntactic and semantic information (e.g. f1, f2, f3, f4, f5). The features can be classified according to their relevance to the linguistic levels. The first 5 features make use of the morphological, syntactic and semantic characteristics of anaphoric relations. As for f1, a subject-antecedent tends to be the most likely candidate for the anaphora resolution. This is reflected in the centering theory in the prominence hierarchy. The underlying assumption of the semantic class determination (concerning f5) is that the semantic class for a zero subject and the antecedent has to be identical.

The feature f6 is a coreference-level feature and it describes the relation between antecedents and zero subjects. McEnery et al.(1997) examined the distance of pronouns and their antecedent and concluded that the antecedents of pronouns do exhibit clear patterns of distribution.

In addition, we introduced a set of features (f7, f8, f9, f10, f11) reflecting the properties of Spanish. In Spanish, possessive adjectives and reflexive pronouns can also give some clues for the person of the antecedent because of their morphological information. Feature f7 and f8 reflect this property. As for f9, there are many extra-sentential zero subjects in Spanish spoken texts, which means they don't have any antecedents. The presence of antecedent could offer information to find zero subjects. Feature f10 and f11 are about the mood and tense of sentence.

To evaluate the ML approach, we built a corpus of 1000 sentences in which a zero subject is included and in which morphological information is not enough to restore the omitted subject.[4] The sources of the corpus were 9 movie scripts and 12 drama episodes.

Among 11 subject types in the corpus, we discarded 4 subject types whose number of frequency is less than 10, as we thought that they belong to rare cases. For this reason, only 988 sentences were tested.

There are 7 subject types in 988 sentences and the number of frequency for the subject types is presented in Table 2.

Subject type	**Frequency**
él(he)	290
ella(she)	276

[3] In Spanish, all nouns are either masculine or feminine. However, the gender of some antecedents such as 'yo(I)' and tú(you)' can vary according to their referents. In these cases, we consider that they have a 'neutral' gender.

[4] 1000 sentences are not large enough to train and to validate the classifier. However, as the building of the training corpus for Spanish zero subject resolution is time consuming and labor-intensive, the experiment was conducted with the corpus of 1000 sentences. The construction of the training corpus is still on-going.

yo(I)	275
tú(you-informal)	53
Usted(you-formal)	43
ellos(they)	32
Ustedes(you-formal plural)	19

Table 2. The number of frequency for the subject types

4.1 Experiment 1

All experiments were performed using 'WEKA' (3.6.10 version). We selected SVM (Support Vector Machine) algorithm. By performing 10-fold cross validation as a test option, the results were obtained.

Using 11 features we proposed, 83.6% for accuracy was reported. For comparison, a simple baseline would be to assume that we randomly choose one subject candidate among three most frequent subject types (él, ella, yo). The accuracy of this method would be about 33.3%. Though it might not be a quite fair comparison, the proposed method could improve the accuracy for about 50% over the baseline.

	baseline	**our method**	**remark**
Accuracy	about 33%	83.6%	about 50% improved

Table 3. The result of experiment 1

Precision, recall and f-measure for each subject type are as follows.

Subject Type	**precision (%)**	**recall (%)**	**f-measure (%)**
tú	0.962	0.943	0.952
yo	0.884	0.971	0.925
él	0.915	0.779	0.842
ella	0.774	0.917	0.839
ellos	0.558	0.906	0.69
Usted	0.2	0.023	0.042
Ustedes	0	0	0

Table 4. precision, recall and f-measure for each subject type

The values of f-measure for the subject types 'tú', 'yo', 'él', 'ella' were higher than the other subject types. We assume that the training instances for 'yo', 'él', 'ella' were relatively enough to be trained by the system (275 for 'yo', 290 for 'él', 276 for 'ella').

On the other hand, the token frequency for the subject type 'tú' was far less than the three subject types above. We assume the reason why the value of F-measure for the subject type 'tú' is the highest as follows. Feature 'f11' has a value which is for imperative sentence and in the corpus about 94.5% of imperative sentence has a subject type 'tú'. If 'f11' is eliminated, the value of F-measure dropped from 0.952% to 0.685%.

The following table shows the ranking of the features selected by using 'InfoGainAttribute Evaluator'.

Ranking	**Feature**
1	f4
2	f2
3	f10
4	f1
5	f6
6	f3
7	f11
8	f5
9	f9
10	f8
11	f7

Table 5. The ranking of 11 features

The feature 'f4' which is about the gender of the antecedent ranked top and then 'f2' which is about the person of the verb ranked second. These features might play an important role to give information about the gender and person of the zero subject.

4.2 Experiment 2

We conducted another experiment to find out the best feature combination for the zero subject resolution. The accuracy is measured by eliminating features from the lowest ranking one by one. Table 6 shows the condition of the experiment and Figure 2 its result.

ID	Condition of the experiment
1	f7 eliminated
2	f7, f8 eliminated
3	f7, f8, f9 eliminated
4	f7, f8, f9, f5 eliminated
5	f7, f8, f9, f5, f11 eliminated
6	f7, f8, f9, f5, f11, f3 eliminated
7	f7, f8, f9, f5, f11, f3, f6 eliminated
8	f7, f8, f9, f5, f11, f3, f6, f1 eliminated
9	f7, f8, f9, f5, f11, f3, f6, f1, f10 eliminated
10	f7, f8, f9, f5, f11, f3, f6, f1, f10, f2 eliminated

Table 6. Condition of experiment 2

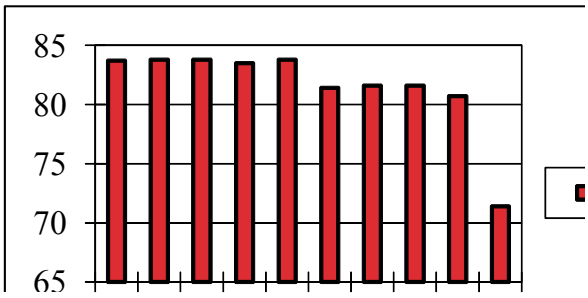

Figure 2. The result of experiment 2

There was very little difference between the accuracy when the 5 low rank features were eliminated and the accuracy when 11 features were used. If the feature 'f3' which is about number of the verb is eliminated, the accuracy decreased about 2%. In other words, the 5 low rank features may be regarded as not significant ones to classify subject types.

5 features that did not have a great influence on classifying subject types are as follows. Feature 'f7' is about presence of a possessive adjective in the same sentence and feature 'f8' is about the presence of a reflexive pronoun in the same sentence. In the corpus, there are 956 and 855 cases where the possessive adjective and the reflexive pronoun don't exist, so because of the occurrence frequency, these features might be of little importance. Feature 'f9' is about presence of antecedent and there are about 41% of sentences which don't have antecedent. Therefore, whether an antecedent exists or not may not be crucial in zero subject resolution. Feature 'f5' is about the semantic class of an antecedent and there are lots of cases where the antecedent doesn't exist as mentioned above, so this feature might also not be significant to classify subject types. 'F11' is the feature about the tense of sentence. Based on the results, the tense of sentence doesn't seem to play a significant role in zero subject resolution.

4.3 Experiment 3

We performed an experiment to identify which features contribute most to the 3 subject types, 'él', 'ella', 'yo', that showed the highest frequency in the corpus. As for the 3 subject types, the f-measure values showed little difference when the 5 lowest rank features were eliminated one by one. So we tried to eliminate the high rank features and compare the f-measure value with the case in which 11 features are used for the zero subject resolution. Table 7 presents the results of the experiment.

	f-measure (%)		
	11 features are used	**f4 eliminated**	**f2 eliminated**
él	0.842	**0.494**	0.85
ella	0.839	**0.443**	0.843
yo	0.925	0.82	**0.691**

Table 7. The result of experiment 3

These results show that as for 3rd person singular subject ‘él’, ‘ella’, when the feature ‘f4’ about gender of antecedent was eliminated, the value of f-measure decreased sharply. Feature ‘f4’ has a value to distinguish between the 3rd person masculine singular and 3rd person feminine singular subject, so it might affect to classify between the 3rd person singular subject ‘él’ and ‘ella’.

In case of ‘yo’, f-measure value decreased sharper when ‘f2’ which is about the person of verb was eliminated than when ‘f4’ was removed. Feature ‘f2’ has a value to distinguish between the verbs which have the same verb ending in case of 1st and 3rd person, so it could be a significant feature to classify ‘yo’ as a right subject type.

5 Conclusion

In this paper, we proposed a hybrid approach to resolve Spanish zero in developing Spanish-to-Korean MT. It combines the linguistic knowledge and ML approach in one model. For the case in which a zero subject couldn’t be resolved using verb ending, the ML method was employed. To utilize ML, 11 features were suggested for Spanish zero subject resolution. In order to identify the feasibility for our method, several experiments were conducted. The accuracy was about 83.6% which was about 50% higher than the baseline when 11 features were used for the ML.

We performed other experiments to find out the best feature combination and the specific feature to classify the subject types which showed high frequency in the corpus. As a result, we figured out 5 features which were not significant for the zero subject resolution and 2 features which played an important role to classify high frequency subject types.

Currently we are increasing the size of the training corpus to balance the various subject types. In the future we are planning to validate our model in depth with the new training corpus.

Acknowledgments

This work was supported by the IT R&D program of MSIP/KEIT. [10041807, Development of Original Software Technology for Automatic Speech Translation with Performance 90% for Tour/International Event focused on Multilingual Expansibility and based on Knowledge Learning]

References

Brennan et al. (1987) A centering approach to pronouns. In Proceedings of the 25th Annual Meeting of the ACL (ACL'87), pp. 155-162.

Carbonell, J.G. and R.D. Brown. (1988) Anaphora resolution: a multi-strategy approach, In Proceedings of the 12. International Conference on Computational Linguistics (COLING'88), Vol.I, pp. 96-101.

Dahl, D.A. and C.N. Ball. (1990) Reference resolution in PUNDIT. Research Report CAITSLS-9004. Paoli: Center for Advanced Information Technology, pp. 168-184.

Ferrández, A. and J. Peral. (2000) A computational approach to zero-pronouns in Spanish. In Proceedings of the 38th Annual Meeting of the Association from Computational Linguistics (ACL'00), Hong Kong, October, pp. 166-172.

Han, N. (2004) Korean null pronouns: classification and annotation. In Proceedings of the Workshop on Discourse Annotation. 42nd Annual Meeting of the ACL-04, pp. 33-40.

McEnery et al. (1997) Corpus annotation and reference resolution. In Proceedings of the ACL Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts, pp. 67-74.

Mitkov, R. (1994) A new approach for tracking center, In Proceedings of the International Conference "New Methods in Language Processing" (NeMLaP-1), pp. 13-16.

Okumura, M. and K. Tamura. (1996) Zero pronoun resolution in Japanese discourse based on centering theory. In Proceedings of the 16th International Conference on Computational Linguistics (COLING'96), Copenhagen (Denmark), pp. 871-876.

Palomar et al. (2001) An Algorithm for Anaphora Resolution in Spanish Texts, Computational Linguistics, Vol. 27, Num. 4, pp. 545-567.

Rello et al. (2010) A machine learning method for identifying non-referential impersonal sentences and zero pronouns in Spanish. Procesamiento del Lenguaje Natural, 45, pp. 281–287.

Rich, E. and S. LuperFoy. (1988) An architecture for anaphora resolution, In Proceedings of the Second Conference on Applied Natural Language Processing (ANLP-2), pp. 18-24.

Sidner, C. (1986) Focusing in the comprehension of definite anaphora, Readings in Natural Language Processing ed. by B. Grosz, K. Jones & B. Webber. Morgan Kaufmann Publishers, pp. 363-394.

Stys, M.E. and S.S. Zemke. (1995) Incorporating discourse aspects in English – Polish MT: towards robust implementation, In Proceedings of the international conference "Recent Advances in Natural Language Processing" (RANLP'95).

Walker et al.(1994) Japanese Discourse and the Process of Centering, Computational Linguistics, Volume 20, Number 2, pp. 193-232.

Wilks, Y. (1973) Preference semantics. Stanford AI Laboratory memo AIM-206. Stanford University.

Zhao, S. and H.T. Ng. (2007) Identification and resolution of Chinese zero pronouns: a machine learning approach. In Proceedings of the 2007 Joint Conference on EMNLP/CNLL-07, pp. 541-550.