Genealogical Record Linkage on International Data

Randy Wilson
FamilySearch.org
Salt Lake City, Utah
randy@axon.cs.byu.edu

Introduction

Record linkage is used to identify multiple records that refer to the same real person in many fields, including medicine, advertising, business and genealogy. Record linkage requires dealing with variations in names, dates, places and other data, which can be a challenge even in a single culture. FamilySearch deals with genealogical records from many parts of the world, which presents further challenges in dealing with various naming conventions, writing systems, and cultural differences. This paper reviews recent work done by FamilySearch to address the special matching challenges on Chinese, Japanese, Korean and Cyrillic records, as well as Scandinavian cultures.

Record Linkage Overview

Record linkage is a term first coined in the medical field (Dunn, 1946), and involves identifying multiple records that refer to the same real entity. Traditional genealogy involves finding information about real people in a variety of records that they appear in. For example, a person may appear in a birth certificate, a county registry, a church christening record, a marriage certificate, several census records, tax and land records, journals, family photos, a death certificate and tombstone. Each appearance of the person in a record can add new information, confirm previously known information, or even bring previous information into question. The work of genealogy is largely involved with locating such data and drawing conclusions based on all of the available evidence. Record linkage can play a valuable role in helping to bring together records that have been put into electronic databases.

Probabilistic record linkage. *Probabilistic Record Linkage* (PRL) (Newcombe et al., 1959; Fillegi & Sunter, 1969; Newcombe, 1988) has been the most popular technique for record linkage in most industries. This method involves calculating a *field agreement weight* and a *field disagreement weight* for each of several *fields*, such as given name, surname, birth date, and so on. This method was used during past decades by the Family History Department in such products as *Ancestral File*and *TempleReady*.

Neural networks and complex features. In recent years, FamilySearch has used more powerful techniques for record linkage, including neural networks for training weights and the use of multi-valued *features* instead of simple field comparisons. For example, instead of using simple field agreement or disagreement on given names, several levels of agreement can be used, ranging from multiple given name pieces matching exactly ("John Henry" vs. "John Henry") to several levels of name similarity ("John Henry" vs. "John H.", "Jonathan" or "Johannes") to conflicting data ("John Henry" vs. "John Albert", "John A." or even "Frederick Simon"). Similarly, several levels of agreement can be used for dates and places. In addition, more complex features can be used that use data from a variety of fields in order to detect situations such as when the person in one record died before the person in the other record was born.

A recent study (Wilson & Quass, 2008) showed a dramatic improvement in accuracy on a hand-labeled set of 80,000 pairs of genealogical records, by using a combination of neural network training and complex features when compared to traditional probabilistic record linkage formulas applied to simple field comparisons.

Data variation. One of the first challenges that must be overcome in a record linkage system is to account for variation in the data. Names have variations due to nicknames ("Bob" vs. "Robert"), married names vs. maiden names ("Elizabeth Turner" vs. "Elizabeth Smith"), spelling variations ("Elizabeth" vs. "Elisabeth"), initials ("John Henry" vs. "John H."), typographical errors, illegible handwriting, and so on.

Dates can vary due to formatting differences ("12 Jun 1850", "6/12/1850", "1850.12.6"), estimates ("1850", "about 1848"), typos ("1701" vs. "1710"), or even calendar changes.

The same place is often spelled differently, abbreviated in various ways, or can be subject to boundary or name changes over time. Sometimes a city in one county or country at one period of time is in another one later on. There are also ecclesiastical and governmental boundaries that overlap in various ways, so sometimes the same place can fit into more than one place hierarchy.

The result of these types of variation is that pairs of records that refer to the same person often appear to disagree when they in fact are saying the same thing. If not corrected, this would cause a classifier to be trained on "noisy" data, making it much less accurate, and would cause errors during classification.

Normalization and Standardization. To account for known types of data variation, then, it is often possible to do *normalization* or *standardization*. In this paper, the term *normalization* refers to things like converting names to lower case, handling punctuation, tokenizing name pieces, etc. The term *standardization* refers to a more extensive conversion of data, such as looking a place up in a place catalog to get a place ID; parsing dates and converting them to a standard day/month/year format; getting a name group ID of a set of names that are regularly interchanged, etc.

These techniques can greatly reduce the variation in the data, making agreement and disagreement much more consistent among matching and differing pairs. This in turn produces a much more accurate classifier.

Special Challenges of International Data

We have recently focused on the problem of matching on records in Chinese, Japanese, Korean and Cyrillic languages, as well as Scandinavian cultures. There are many things that remain the same across cultures and languages, because we are all part of the same human family. People everywhere tend to have names; birth, death and marriage dates and places. They also tend to have fathers, mothers, spouses and children. However, there are some unique challenges in each of these cultures that need to be handled in order to apply record linkage properly.

Scripts. One of the first challenges with matching in some languages is the variety of scripts that are used. Unicode has made dealing with these scripts much easier than it once was, but there is still the common problem that several languages use more than one script as well as "romanized" versions of names or places.

Chinese names and places are typically written using Chinese characters, of which there are many thousands in common usage.

Japanese names use three main script types: *kanji*, which uses essentially the same characters as Chinese (i.e., the same Unicode characters); and two phonetic scripts, *hiragana* and *katakana*. Since these two phonetic scripts have a one-to-one mapping, it is possible to map one to the other during normalization for matching purposes (though it is of course wise to preserve the original for display and canonical storage).

Korean names use two script types: *hanja*, which again uses essentially the same characters as Chinese; and a phonetic script, *hangul*. Hangul is written in clusters of letters, each cluster constituting a syllable. Each syllable of a name typically corresponds to a single hanja (Chinese) character. There are some names that are pronounced the same—and thus appear the same in hangul—but which have different hanja characters. Thus, hanja can be slightly more distinguishing than hangul, although hangul is usually the preferred name form for everyday use.

Cyrillic is a phonetic script used in Russia and other nearby countries.

One thing common with all four of these cases is that names can be *romanized*, meaning that they can be transliterated into a latin script. In the case of Japanese and Korean, there are also multiple scripts in the language itself that the name can appear in.

Thus, in dealing with these scripts, we often have to deal with a single name having several *name forms*. During matching, it makes little sense to compare name forms of different scripts (e.g., kanji and katakana; or Chinese and romanized spellings), since they cannot agree without transformation. The method of romanization also tends to vary over time, so romanization is often inconsistent. Since spelling variation is rare in Chinese, Japanese and Korean, we have found it useful to ignore the romanized forms of names when both records being compared have the same non-roman script. For example, if two Korean records both had hangul forms for the names, we could ignore the romanized version of the name. This is actually important to do, because sometimes the romanized names look very similar, and would thus appear as "nearly agreed", when in fact they firmly conflict in hangul.

Name order. In western cultures, the given name usually comes first and the surname appears at the end, which is why it is often called the "last name." In most Asian cultures, however, the surname comes first, and the given name comes last. Some computer systems have made it awkward to use the natural name ordering, so in matching we have to be on the lookout for cases where surnames were given last (e.g., when names were entered using a western template).

Spaces as delimeters. Names in western cultures tend to have spaces between the name pieces. "John Albert Smith" would look odd if ran together as "johnalbertsmith" or even "johnalbert smith". In Asian names and places, however, spaces are less crucial, and are often not used.

"Mrs." When matching names in English and similar languages, we recognize some words like "Mr.", "Mrs.", "Dr.", "infant", etc., as not being real names. We also recognize that often we get a husband's name listed in a wife's name fields. For example, when comparing "Mrs. John Smith" against "Elizabeth Turner", we might initially assume that these names conflict. But if both of these people are married to a "John Smith", then in the first case we would drop "Mrs." as a "noise word", and drop both "John" and "Smith" as being copied from her husband. Then we would accurately say that we do not know the first person's name, and so although we can't say that the names of the two records agree, neither do we apply a penalty for conflicting. (Meanwhile, we will also end up comparing their spouses' names in a separate feature of the classifier).

The same sort of thing happens in Chinese,Japanese, Korean and Cyrillic. In Chinese, a wife's name is often given as the husband's name with "fu-ren" (夫人) after it, which means "Mrs." or "wife". Since there is usually no space between the name and these characters, they have to be detected at the end of a name. As an example, consider the first row in Table 1. Even if you can't read Chinese, you can probably recognize that the first three characters are the same in the husband's and wife's name, and that the wife has the "Mrs." characters (夫人) at the end.

Language	Husband's name	Wife's name
1. English Mrs. [<given>] <surname>	John Smith	Mrs. Smith; or Mrs. John Smith
2. Chinese	黃 德纘 (Huang Te-Tsuan)	黃 德纘夫人 (Huang Te-Tsuan; Mrs.)
3. Japanese <surname>夫人	鈴木 栄吉 (Suzuki Eikichi)	鈴木夫人 (Mrs. Suzuki)
4. Japanese <surname>夫人<given>	鈴木 栄吉 (Suzuki; Eikichi)	鈴木夫人 かの (Suzuki; Mrs.; Kano)
5. Korean <surname given>의 부인	김 성수 (Kim Seong-Su)	김 성수의 부인 (Kim Seong-Su's wife)
6. Cyrillic	Иван Овсянников Ivan Ovsyannikov	Госпожа Ивана Овсянникова or: Г-жа Ивана Овсянникова (Mrs. Ivana Ovsyannikova)

Table 1. Examples of "Mrs." in English, Chinese, Japanese, Korean and Cyrillic.

In Japanese, the same Chinese characters (夫人) are also often used after a husband's name in the wife's name. However, there are two common patterns that occur commonly. The first is to append these two characters after the husband's surname, i.e., <surname>夫人. The second is to additionally include the wife's given name, if it is known, i.e., <surname>夫人<given>. In the latter case, the surname is often in "kanji" (Chinese characters), but the given name is typically in katakana or hiragana, as shown in Row 3 of Table 1.

In Korean, the suffix "ui buin" (hangul: "의 부인") is commonly used to indicate "wife of". The "ui" ("의") part is the posessive, like the English "of" or "'s"; and "buin" ("부인") means "wife". Usually the posessive is attached directly to the husband's name, and there is often—but not always—a space before the "wife" part.

In Cyrillic, one word used for "Mrs." (or "Lady") is "gospozha" ("Госпожа", also abbreviated "Г-жа"). This word comes at the beginning of the wife's name, like "Mrs." does in English. When romanized, the translation of this word is often given as "Frau" or "Frau des" (German for "wife of"). One unique thing that happens in Russian and similar cultures is that the female version of a name tends to end in "a", even when the husband's name is used in the wife's name. So, as shown in Table 1, row 6, the wife of Ivan Ovsyannikov is written as "Mrs. Ivana Ovsyannikova", at least in Cyrillic.

Daughter/son. In both Chinese and Korean, it is common to find the symbol “ssi” (“氏” in Chinese or in hanja; “씨” in hangul) after a surname. This means something like “Mr.” or “Miss”, indicating that we do not know the person’s given name, but do know their surname. Often, for example, the female’s given name is not listed in a record, but her surname is, so this symbol keeps the name from looking awkward. Table 2 shows an example in which the hanja (Chinese) character “氏” follows the hanja version of the surname “Kim” (“金”); the hangul “ssi” (“씨”) follows the hangul version of “Kim” (“김”), and the Roman transliteration as well as English translations are also shown. A Chinese example would skip the hangul version.

Hanja:	金 氏
Hangul:	김 씨
Romanized:	Kim ssi
English:	“Miss Kim”

Table 2. Korean forms of “Miss Kim.”

In Japanese, a different convention appears in our data. A female with a known maiden name but an unknown given name is sometimes listed using the symbol “娘”, meaning “daughter” or “miss”. So a female with the surname “Suzuki” (“鈴木”) could be listed as “Miss Suzuki” (“鈴木娘”; literally “Suzuki-daughter”).

Similarly, a son can have the symbols for “son” (“息子”) appended to their surname.

By recognizing these patterns in the names, we can avoid saying that names conflict when they really do not, and, perhaps more importantly, we can avoid saying that many, many records agree, when really they are just all saying “Mr.”, “Mrs.” or “Miss”.

Patronymic names. In Scandinavian countries such as Sweden, Norway, Denmark, Iceland, it is common in genealogical records to find *patronymic* naming patterns, in which a person uses their father’s name as part of one of their names. For example, someone named Olaf whose father’s name was Sven could have the name “Olaf Svensen”, meaning “Olaf, the son of Sven.” If Olaf then has a daughter named Inga, her name would be listed as “Inga Olafsdotter” (rather than “Inga Svensen”), i.e., her last name comes from her father’s first name rather than from his last name as in other western countries. There are many spelling variations of “son” (son, sen, ssen, etc.) and “daughter” (dotter, dtr, etc.) that need to be handled by the matching algorithm.

In later years—or when migrating to another country—some people started using their father’s patronymic name as their surname. For example, Inga, daughter of Olaf Svensen might go by the name Inga Svensen if she migrated to America. When matching these types of records, it is not uncommon to get two records about the same person with both naming types (e..g, “Inga Olafsdtr” vs. “Inga Svensen”, both listed as the daughter of Olaf Svensen).

Matching names in these cultures requires normalization of the Scandinavian name stem (e.g., convert “son”, “sen”, “dtr”, “dotter”, etc., to “son” for the purpose of looking for “near agreement” between surnames, if an exact agreement is not found). It also can benefit from considering the father’s name when determining if names conflict or not.

Cyrillic patronymic names. In Russia and eastern European cultures, it is common for a person to take their father’s given name as a middle name and their father’s surname as a surname.

The patronymic middle name typically ends with an ending like "-vich" () for males, and "-yevna", "-ovna" or "-ichna" for females. For example, a male named Sergey (Сергей) who is the son of Ivan Popov (Иван Попов) would typically have the name Sergey Ivanovich Popov (Сергей Иванович Попов). Similarly, a daughter of Ivan Popov named Tatjana (Татьяна) would have the name Tatjana Ivanova Popova (Татьяна ИвановнаПопова).

Parsing Asian places. Places written in Chinese (or Korean hanja or Japanese kanji) tend to be written from general (e.g., country or province) to specific (e.g., town or village), often with no spaces or other delimeters. In order to parse these places, we found it useful to start at the leftmost character and look each substring up in our place catalog (i.e., length 1, length 2, etc.). When a matching entry was found, the remaining part of the string was then parsed the same way, except that places were only recognized if they appeared in the catalog *and had the earlier places above them in the place hierarchy*.

For example, in parsing the place "中國廣東省", we would find that the first two characters match the place "China", and that the next three then match the place "Guangdong, China." If there were a "Guangdong, Taiwan," too, we would ignore that interpretation, since the first two characters restrict our search to only the places within China.

At first glance it might appear that this technique was in danger of exploding exponentially, since we do not do a "greedy" search, but continue looking for additional matches in the catalog even after finding a substring. (For example, in the above case, we would look for a 3-, 4- and 5-character place name even though we found "China" in the first two characters). However, very few substrings actually match a place in the catalog, and fewer still match a place that is within a higher-level place found earlier in the string, so it would be difficult to construct cases where more than a few lookups are done.

Results and conclusions

Before giving serious focus to non-Roman matching, FamilySearch was able to do matching that was still useful, but suffered from cases where there were many irrelevant matches being displayed to users and other cases where good matches were not appearing.

In one study of place standardization, none of the Asian places without delimeters was recognized correctly, whereas most of them were recognized at least as deep as our catalogs went as soon as we began using the left-to-right parsing and searching technique.

In another study of 5,000 pairs of possible duplicate Korean records, almost half were affected by recognizing the "Mrs." ("wife of") or "Mr./Miss" (unknown given name) cases. In a set of 4,700 Japanese pairs of records, about 12% of the pairs had their name comparisons affected by handling "Mrs." properly, and in a set of 877 pairs of Chinese records, 22% of name comparisons were affected.

Before handling those key words properly, some records had hundreds of bogus potential matches listed, even though it was clear to users that these were not the same person.

There are certainly additional subtleties that will surely help to improve matching accuracy on non-Roman records further. However, the cases reviewed in this paper were so common that they tended to mask other remaining issues. With these cases handled much more reasonably, we hope to be able to continue to progress in improving accuracy on these and other cultures.

References

Dunn, H. L. (1946). Record Linkage, *American Journal of Public Health*, **36**, 1412-1416.

Fillegi, I. P. and Sunter, A. B. (1969). A Theory for Record Linkage, *Journal of the American Statistical Association*, **64**, 1183-1210.

Newcombe, H. B., Kennedy, J. M., Axford, S.J., and James, A.P. (1959). Automatic linkage of vital records. *Science*, **130**, 954-959.

Newcombe, Howard B. (1988). *Handbook of Record Linkage*, Oxford University Press, New York.

Wilson, Randy, and Dallan H. Quass, (2008). "Beyond Probabilistic Record Linkage: Using Neural Networks to Improve Genealogical Record Linkage," submitted to *The Twenty-third Conference on Artificial Intelligence (AAAI'08).*